Ricardo César Gonçalves Sant'Ana
Moisés Lima Dutra
Guilherme Ataíde Dias
Organizadores

ANAIS
WIDaT 2018

UNIVERSIDADE FEDERAL DA PARAÍBA

Reitora MARGARETH DE FÁTIMA FORMIGA MELO DINIZ

Vice-Reitora BERNARDINA MARIA JUVENAL FREIRE DE OLIVEIRA

EDITORA DA UFPB

Diretora IZABEL FRANÇA DE LIMA

Supervisão de Administração GEISA FABIANE FERREIRA CAVALCANTE

Supervisão de Editoração ALMIR CORREIA DE VASCONCELLOS JÚNIOR

Supervisão de Produção JOSÉ AUGUSTO DOS SANTOS FILHO

Capa: André Luiz Dias de França

Título: Um facho de luz através da janela do conhecimento

Catalogação na Publicação
Seção de Catalogação e Classificação

W926a Workshop de informação, dados e tecnologia (2 : 2018 : João Pessoa-PB).
Anais WIDaT 2018, de 27 a 29 de novembro de 2018 / Organizadores : Ricardo César Gonçalves Sant'Ana, Moisés Lima Dutra, Guilherme Ataíde Dias. - João Pessoa : Editora UFPB, 2018.
368 p. : il.

ISBN: 978-85-237-1381-2

1. Ciência da informação. 2. Dados. 3. Tecnologia - Informação. I. Sant'Ana, Ricardo César Gonçalves. II. Dutra, Moisés Lima. III. Dias, Guilherme Ataíde. IV. Título.

UFPB/C CDU: 02

Os artigos e suas revisões são de responsabilidade dos autores.

EDITORA DA UFPB Cidade Universitária, Campus I –s/n

João Pessoa – PB

CEP 58.051-970

editora.ufpb.br

editora@ufpb.br

Fone: (83) 3216.7147

Ricardo César Gonçalves Sant'Ana
Moisés Lima Dutra
Guilherme Ataíde Dias
Organizadores

ANAIS WIDaT 2018

II Workshop de Informação, Dados e Tecnologia

Editora UFPB
João Pessoa
2018

Organização do WIDaT 2018

- **Organização Geral:**

  Guilherme Ataíde Dias (PPGCI-UFPB) - Coordenador geral do evento

  Moisés Lima Dutra (PPGCIN-UFSC) - Vice-coordenador

- **Coordenador da Comissão Científica:**

  Ricardo César Gonçalves Sant'Ana (PPGCI-UNESP)

- **Comissão científica**

  Adilson Luiz Pinto (PPGCIN-UFSC)

  Ana Alice Baptista (Universidade do Minho, Portugal)

  Ana Carolina Simionato (PPGCI-UFSCar)

  Angela Maria Grossi de Carvalho (PPGCI-UNESP)

  Bernardina Maria Juvenal Freire de Oliveira (PPGCI-UFPB)

  Cristian Berrío-Zapata (PPGCI-UFPA)

  Dalton Lopes Martins (FCI-UnB)

  Denysson Axel Ribeiro Mota (PPGB-UFCA)

  Douglas Dyllon Jeronimo de Macedo (PPGCIN-UFSC)

  Ed Porto Bezerra (PPGI-UFPB)

  Edgar Bisset Alvarez (PPGCIN-UFSC)

  Edna Gusmão de Goés Brennand (MPGOA-UFPB)

  Edna Gomes Pinheiro (DCI-UFPB)

  Elaine Parra Affonso (FATEC-SP)

  Elvis Fusco (UNIVEM-Marília)

  Enrique Muriel Torrado (PPGCIN-UFSC)

  Evandro de Barros Costa (IC-UFAL)

  Fábio Paraguaçu (IC-UFAL)

  Fernando de Assis Rodrigues (PPGCI-UNESP)

  Gustavo Medeiros de Araújo (PPGCIN-UFSC)

  Henry Pôncio Cruz de Oliveira (PPGCI-UFPB)

  Joana Coeli Ribeiro Garcia (PPGCI-UFPB)

  José Eduardo Santarém Segundo (USP-FFCLRP)

  Leonardo Castro Botega (UNIVEM-Marília)

  Luana Farias Sales Marques (PPGCI-IBICT-UFRJ)

  Marckson Roberto Ferreira de Sousa (PPGCI-UFPB)

  Luís Fernando Sayão (CNEN)

  Marcelo Morandini (EACH-USP)

  Márcio Matias (PPGCIN-UFSC)

  Marcos Mucheroni (CBD-USP)

  Marynice de Medeiros Matos Autran (PPGCI-UFPB)

Maurício Barcellos Almeida (PPGGOC-UFMG)
Moisés Lima Dutra (PPGCIN-UFSC)
Plácida Leopoldina V. da Costa Santos (PPGCI-UNESP)
Pedro Luiz Pizzigatti Corrêa (POLI-USP)
Renata Baracho (PPGGOC-UFMG)
Ricardo César Gonçalves Sant'Ana (PPGCI-UNESP)
Robson Rodrigues Lemos (UFSC-Araranguá)
Rogério Ramalho (PPGCI-UFSCar)
Ryan Ribeiro de Azevedo (UFRPE-UAG)
Sandra de Albuquerque Siebra (PPGCI-UFPE)
Sandro Rautenberg (DECOMP-UNICENTRO)
Silvana Aparecida Borsetti G. Vidotti (PPGCI-UNESP)
Virginia Bentes Pinto (PPGCI-UFC)
Wagner Junqueira de Araújo (PPGCI-UFPB)
Zaira Regina Zafalon (PPGCI-UFSCar)

- **Coordenador do Cerimonial:**
  André Luiz Dias de França (PPGCI-UFPB)

- **Coordenador da Equipe Técnica Local:**
  Laerte Pereira da Silva Júnior (CCHLA-UFPB)

- **Equipe Técnica Local:**
  Adriana Alves Rodrigues (PPGCI-UFPB)
  Antonio Felipe dos Santos (MPGOA-UFPB)
  Débora Gomes de Araújo (PPGCI-UFPB)
  Pedro Augusto de Lima Barroso (PPGCI-UFPB)
  Pollianna Marys de Souza e Silva (PPGCI-UFPB)
  Renata Lemos dos Anjos (PPGCI-UFPB)

SUMÁRIO

A CIÊNCIA DA INFORMAÇÃO E A EDUCAÇÃO EM GESTAO E CURADORIA DE DADOS DE PESQUISA 13
Gracielle Mendonça Rodrigues Gomes
Marcello Mundim Rodrigues

A CIÊNCIA DE DADOS À LUZ DA CIÊNCIA DO SENSO COMUM ARISTOTÉLICO: uma análise comparativa com a tecnologia sensorial 21
Tereza Cristina Fernandes
Marcello Mundim Rodrigues

A CONSTRUÇÃO DO REPOSITÓRIO DE DADOS DA UFPB: a experiência com o dataset de Arboviroses 31
Pollianna Marys de Souza e Silva
Sandra de Albuquerque Siebra

A DISSEMINAÇÃO DE INFORMAÇÕES ASSOCIADAS À EPIDEMIA DO ZIKA VÍRUS: Uma pesquisa a partir de dados coletados no Twitter 37
Guilherme Ataíde Dias
Denysson Axel Ribeiro Mota
André Luiz Dias de França
Eduardo de Santana Medeiros Alexandre

ANÁLISE DA RELAÇÃO ENTRE PERFIL E DESEMPENHO ACADÊMICO DOS ALUNOS MATRICULADOS NA DISCIPLINA DE INTRODUÇÃO À PROGRAMAÇÃO UTILIZANDO ALGORITMOS DE CLASSIFICAÇÃO 43
Roberto Silva de Oliveira Júnior
Douglas Valentim de Almeida Cardins
Wenddell Juler Pereira de Lima
Thereza Patrícia Pereira Padilha
Vanessa Farias Dantas

ANÁLISE DE SENTIMENTOS: Identificando sentimentos em comentários da Rede Humaniza SUS 50
Eduardo Alves Silva
Luis Felipe Rosa de Oliveira

ARBOVIS: Uma aplicação de visualização de dados sobre Arboviroses 57
Talitha Alessandra Medeiros Ribeiro
Larissa de Sousa Moreira Gusmão
Damires Yluska de Souza Fernandes

AS LICENÇAS CREATIVE COMMONS E SOFTWARE LIVRE COMO FORMAS DE PROTEÇÃO NA SOCIEDADE DA INFORMAÇÃO COMPARTILHADA 68
Maria Carolina Zanini Ferreira
Clarissa Stefani Teixeira
Enrique Muriel-Torrado

ASPECTOS DO CICLO DE VIDA DE DADOS EM PROCESSOS DE CONSTRUÇÃO DE ONTOLOGIAS BIOMÉDICAS 76
Jeanne Louize Emygdio
Eduardo Ribeiro Felipe

CIÊNCIA ABERTA:
os desafios para concepção da ciência cidadã 89
Luciana Gonçalves Silva Souza
Izabel Antonina de Araújo
Júlia Rocha Ribeiro

DESAFIOS TECNOLÓGICOS NO PROCESSO DE RECEPÇÃO NORMATIVA: adequação e conformidade na perspectiva da Lei de Proteção de Dados Pessoais 97
Andrea Carmo Name Willemin
Geralda Magella de Faria
Cláudio José Amante

DESIGN AUDIOVISUAL PARA SISTEMAS GENERATIVOS:
Um experimento de construção audiovisual personalizada 109
Rafael M. Toscano
Valdecir Becker
Helder Bruno A. M. de Souza
Marianna C. Teixeira

DIREITO AUTORAL + CREATIVE COMMONS: uma solução para a academia na Sociedade da Informação? 118
Enrique Muriel-Torrado
Rebeca Silva Fernandes de Moura

DIREITO DE PRIVACIDADE NO BRASIL:
Tendências 125
Rossana Gleucy de Ávila Chagas e Carvalho
Márcio Matias

EVASÃO ESTUDANTIL E CIÊNCIA DE DADOS:
primeiros passos de uma pesquisa aplicada no contexto da Educação a Distância da Universidade Estadual do Centro-Oeste 132
Sandro Rautenberg
Alan H. Costa
Paulo R. V. do Carmo
Renan A. M. Nutse
Maria A. C. Knüppel
Marta C. R. Anciutti

GESTÃO DE DADOS DA BIODIVERSIDADE:
aplicação do padrão de metadados Darwin Core 141
Filipi Miranda Soares
Raíssa Yuri Hamanaka
Benildes C. Moreira dos Santos Maculan

HUMANIDADES DIGITAIS:
visualização da produção científica 147
Mirelys Puerta Díaz
Edgar Bisset Alvarez
Silvana Aparecida Borsetti Gregorio Vidotti

IDENTIFICAÇÃO DE ENTIDADES DESTAQUE PARA A MELHORIA DA VISUALIZAÇÃO NA ANÁLISE DE GRAFOS 157
Roberto Zaina
Vinícius Faria Ramos Culmant
Gustavo Medeiros de Araujo

INFORMAÇÃO E PRIVACIDADE NA LEI GERAL DE PROTEÇÃO DE DADOS 175
Rosilene Paiva Marinho de Sousa
Fernando Antônio de Vasconcelos
Marckson Roberto Ferreira de Sousa

LETREIROS E AMORES:
uma estratégia informacional para existir no Instagram 183
Matheus José Pessoa de Andrade
Tassyara Onofre de Oliveira

METADADOS E TWITTER:
uso do Tweet Object para identificação de local 193
Denysson Axel Ribeiro Mota
Gracy Kelli Martins

METAS, AÇÕES E INDICADORES COMO SUBSÍDIOS PARA ANÁLISE DA QUALIDADE DE DADOS:
Uma inversão necessária entre consequentes e antecedentes 200
Pedro Jácome de Moura Jr
Maicon Henrique Ferreira Aragão

MÉTODO MULTICRITÉRIOS PARA AUXILIAR O PROCESSO DE TOMADA DE DECISÃO SOBRE EMPREENDIMENTOS IMOBILIÁRIOS UTILIZANDO MODELOS BIM 207
Bruno Cesarino Soares
Rogério Amaral Bonatti
Renata Maria Abrantes Baracho

NARRATIVAS DE VISUALIZAÇÕES GUIADAS POR DADOS:
Big Data e inovação 217
Adriana Alves Rodrigues

O ENSINO DA E-SCIENCE NO ÂMBITO DOS PROGRAMAS BRASILEIROS DE PÓS-GRADUAÇAO EM CIÊNCIA DA INFORMAÇAO 226
Renata Lemos dos Anjos
Débora Gomes de Araújo

O PAPEL DO BIBLIOTECÁRIO NO PROCESSO DE RECUPERAÇÃO DA INFORMAÇÃO 233
Miriam Cândida de Jesus
Nina Cláudia Mendonça Campos de Miranda
Célia da Consolação Dias

O PAPEL DOS METADADOS NA CURADORIA DIGITAL 241
Rachel Cristina Vesu Alves

O TERMO BIG DATA:
quebra de paradigma dos n-V's 246
Marcos de Souza
Fernanda Gomes Almeida
Renato Rocha Souza

OS PRINCÍPIOS FAIR COMO DIRETRIZES PARA MAXIMIZAR O USO E (RE)USO DOS DADOS DE PESQUISA:
traçando um paralelo com a gestão de documentos arquivísticos digitais 254
Bárbara Carvalho Diniz
Guilherme Ataíde Dias
Alzira Karla Araújo da Silva

PADRÕES DE METADADOS PARA DESCRIÇÃO DE DADOS:
panorama dos repositórios de dados na América Latina 262
Felipe Augusto Arakaki
Ana Carolina Simionato
Paula Regina Ventura Amorim Gonçalez
Plácida Leopoldina Ventura Amorim da Costa Santos

POLÍTICA DE COMPARTILHAMENTO DE DADOS CIENTÍFICOS:
a adoção nos periódicos da Ciência da Informação 269
Renata Lemos dos Anjos
Marynice de Medeiros Matos Autran

PORNOGRAFIA INFANTIL:
análise de manuais de pedofilia 277
Marcelo da Silva Morreira
Enrique Muriel Torrado

PRÁTICAS DE GESTÃO DE DADOS:
uma Revisão da Literatura Sobre o Termo Data Life Cycle *285*
Débora Gomes de Araújo

PREVISÕES DE AUSÊNCIA EM ATENDIMENTO ESPECIALIZADO EM UNIDADES DE SAÚDE PÚBLICA DE FLORIANÓPOLIS 293
Milena Teixeira
Leandro Pereira Garcia
Ingo Ramos
Lucas Alexandre Pedebôs
Douglas Macedo
Gustavo Medeiros de Araujo

PROTEÇÃO AUTORAL DE DADOS:
Uso de Blockchain 304
Elizabete Cristina de Souza Aguiar Monteiro
Jacquelin Teresa Camperos Reyes
Elaine Parra Affonso
Ricardo César Gonçalves Sant'Ana

RECONHECIMENTO DE ENTIDADES NOMEADAS EM RELATÓRIOS DE INTELIGÊNCIA FINANCEIRA 312
Jairo Santana
Diefferson K. Môro
Rogério de Aquino Silva
Vinicius F. C. Ramos
Gustavo Medeiros de Araújo

REPOSITÓRIOS DE DADOS DE PESQUISA:
investigando sua adoção nos institutos brasileiros de pesquisa 319
André Luiz França de Madeiro

REVOLUÇÃO CIENTÍFICA?
Reflexões para uma teoria crítica da ciência dos dados 325
Max Melquiades da Silva
Simone Cristina Dufloth

SUPORTE À GESTÃO DE DADOS DE PESQUISA NAS BIBLIOTECAS:
identificação de serviços a partir da literatura 333
Fernanda Gomes Almeida
Marcos de Souza

TECNOLOGIA BLOCKCHAIN:
um Novo Paradigma no Ciclo de Vida dos Dados 344
Jailton Cardoso da Cruz

WEB SCRAPING DO RESEARCHERID:
proposta de sistema para o monitoramento de Índice H de pesquisadores no Brasil 353
Alexandre Ribas Semeler
Adilson Luiz Pinto
Arthur Oliveira

APRESENTAÇÃO DO WIDAT 2018

Nos dias 27, 28 e 29 de novembro de 2018, na cidade de João Pessoa (PB), mais especificamente no Auditório da Superintendência de Tecnologia da Informação (STI) da Universidade Federal da Paraíba (UFPB), foi realizado o II Workshop de Informação, Dados e Tecnologia (WIDaT 2018), dando sequência às discussões iniciadas no I WIDaT, que ocorreu em setembro de 2017 na Universidade Federal de Santa Catarina (UFSC), em Florianópolis (SC). O WIDaT 2018 abordou temas referentes tanto aos fundamentos teóricos como a aspectos aplicados e sociais que permeiam os estudos relacionados ao acesso, tratamento, armazenamento e uso de dados.

Organizado por pelos professores doutores Guilherme Ataíde Dias (UFPB), Moisés Lima Dutra (UFSC) e Ricardo César Gonçalves Sant'Ana (UNESP), a segunda edição do evento contou com palestras e discussões, transmitidas ao vivo (*on-line*), abordando pesquisas e ações da comunidade acadêmica e profissional, unindo pessoas oriundas de diversas áreas, como Ciência da Informação, Ciência da Computação, Engenharias, Estatística, Matemática, Direito e Administração, entre outras, visando a construção de um ambiente que promovesse amplo diálogo e discussões que possibilitem operacionalizar uma Rede de Pesquisa Brasileira de Acesso a Dados.

O WIDaT 2018 contou com palestrantes de altíssimo nível e que atuam em instituições tanto públicas como privadas. Estas palestras brindaram aos ouvintes com uma ampla percepção das atividades realizadas em prol da Ciência dos Dados, assim como as dificuldades historicamente encontradas no que se refere à atuação profissional nesta área.

As discussões posteriores às apresentações permitiram a interação dos ouvintes, presenciais ou *on-line*, com os palestrantes. O fato dos palestrantes serem pesquisadores de diversas universidades, nacionais e internacionais, assim como representantes do setor privado e do governo, trouxe ao evento uma visão plural da situação da Ciência dos Dados sob perspectivas diferentes, mas complementares, das mesmas dificuldades que surgem nos tempos atuais, e dos caminhos que cada um tem seguido para contorná-las.

Nestes anais trazemos os trabalhos que foram submetidos e julgados, por membros da comissão científica, como pertinentes para as temáticas abordadas ao longo dos três dias. As avaliações foram realizadas por doutores e pesquisadores de diversas instituições e universidades, que contribuíram para a qualidade do WIDaT 2018, não apenas selecionando os trabalhos, mas ao mesmo tempo sugerindo modificações, quando cabíveis, para que estes trabalhos pudessem melhor apresentar seus resultados para a comunidade.

Os trabalhos aqui reunidos abrangem discussões amplas, trazendo desde questões teóricas, voltadas para as demandas epistemológicas e filosóficas da Ciência dos Dados, como elementos práticos sobre a manipulação de dados, tanto para a pesquisa científica como para o governo ou setor privado. Cada capítulo colabora de forma positiva para o crescimento da ciência, com visões abrangentes sobre as problemáticas existentes e possíveis soluções, apresentando também pesquisas realizadas e os resultados obtidos. O contato dos autores encontra-se disponível, estendendo a possibilidade de contato para os interessados no assunto.

A discussão de alguns destes trabalhos durante a realização do evento, possibilitada mediante a organização de mesas de debates após a apresentação presencial dos autores, enriqueceu de grande forma o evento, pois permitiu a interação entre os pesquisadores, a partir de sugestões sobre

novas possibilidades de investigação, perspectivas futuras e possíveis resoluções para os problemas relatados nos textos e apresentações. Acreditamos que estas exposições e discussões enriqueceram não apenas as pesquisas apresentadas, mas toda a comunidade que participou, presencialmente ou à distância, e que teve a oportunidade de dialogar diretamente com os palestrantes e dirimir eventuais dúvidas.

Agradecemos a todos os palestrantes e ouvintes, presenciais e *on-line*, pela participação, que trouxeram discussões sobre questões teóricas e metodológicas de pesquisa bem como suas aplicações práticas. Nossos agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), ao Ministério Público do Estado da Paraíba (MPPB), ao Programa de Pós-Graduação em Ciência da Informação da UFPB (PPGCI/UFPB) e ao Programa de Pós-Graduação em Gestão nas Organizações Aprendentes (MPGOA/UFPB), que colaboraram de forma significativa para a execução do evento e presença dos palestrantes.

E, de forma especial, agradecemos à Reitoria da Universidade Federal da Paraíba (UFPB) que, na pessoa da Vice-Reitora, Profª Drª Bernardina Maria Juvenal Freire de Oliveira, forneceu todo apoio necessário para a realização do WIDaT 2018, participou das sessões de abertura e encerramento, e abrilhantou o evento com discussões extremamente pertinentes para a situação do país. Estes agradecimentos são estendidos também à Superintendência de Tecnologia da Informação (STI), da UFPB, por ter cedido o espaço para realização das sessões e proporcionado apoio operacional, para resolução de pequenos problemas no dia-a-dia, e tecnológico, que possibilitou as transmissões das sessões em tempo real.

Acreditamos que este material reflete o sucesso do WIDaT 2018, resultado do esforço, empenho e dedicação de todos os membros da organização. Desejamos a todas e todos uma ótima leitura.

Denysson A. R. Mota
Universidade Federal do Cariri – UFCA

COMUNICAÇÕES CIENTÍFICAS

A CIÊNCIA DA INFORMAÇÃO E A EDUCAÇÃO EM GESTÃO E CURADORIA DE DADOS DE PESQUISA

LIBRARY AND INFORMATION SCIENCE AND THE EDUCATION ON RESEARCH DATA MANAGEMENT AND CURATION

Gracielle Mendonça Rodrigues Gomes[1], Marcello Mundim Rodrigues[2]
(1) Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, graciellemendonca@yahoo.com.br
(2) Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, marcellomundim@ufu.br

Resumo:
A pesquisa tem como objetivo geral identificar se existem estudos na literatura nacional da CI que contemplem a formação dos profissionais da informação para atuação na área dos dados de pesquisa bem como se existem estudos que tratem sobre desenvolvimento da competência de dados de pesquisa em estudantes e pesquisadores. A busca foi realizada em bases de dados nacionais e internacionais, e no Google Acadêmico. A interpretação dos documentos recuperados foi realizada através da técnica de análise de conteúdo. Não foi identificado nenhum artigo na literatura nacional de CI com as temáticas principais sobre a formação de profissionais da informação na área de dados de pesquisa e/ou o desenvolvimento de competência em dados de pesquisa pelas bibliotecas universitárias e serviços de informação científica. Foi observada a necessidade de discussões e pesquisas sobre essas temáticas na literatura científica nacional de modo a contribuir para a realização de mudanças nos atuais currículos das Escolas de Ciência da Informação no Brasil e o aperfeiçoamento dos serviços de informação prestados aos estudantes e pesquisadores com o intuito de acompanhar as últimas transformações ocorridas na pesquisa científica.
Palavras-chave: Ciência da Informação – Educação. Especialistas em dados de pesquisa. Competência em dados de pesquisa. Gestão de dados de pesquisa. Curadoria digital.

Abstract:
The research has as main objective identifying the existence of studies in the Library and Information Science (LIS) national literature that discuss the education of information professionals to perform in the research data field as well as identifying studies that deal with the development of research data literacy by students and researchers. The search conducted used national and international databases, also Google Scholar. Data analysis and interpretation were performed using the content analysis technique. The research did not identify papers in the national literature of LIS that seek to discuss in their main objectives the education of information professionals in the research data field and/or the development of research data literacy by academic libraries and scientific information services. It observed the need for more discussions and research on these topics in the national scientific literature to contribute to changes in the current curricula of LIS Schools in Brazil and to improve the information services provided to students and researchers in order to follow the latest transformations on scientific research.
Keywords: Information Science – Education. Research data specialist. Research data literacy. Research data management. Digital curation.

1 INTRODUÇÃO

A Ciência da Informação (CI) está em condições de sustentar as mudanças necessárias para favorecer o movimento em direção à área de dados de pesquisa e atender as novas demandas que estão surgindo no contexto da ciência (e-Science). Os problemas enfrentados pelas instituições na gestão e curadoria de dados de pesquisa oferecem uma oportunidade relevante e atraente para a área e seus profissionais reformularem teorias, valores e currículos, redefinirem seus papéis e competências no apoio à pesquisa, além de avançarem na disponibilização de novos serviços, desde que percebam as várias oportunidades que surgirão a partir da integração dos dados em seus serviços e estejam convencidos do valor que podem trazer à pesquisa na medida em que investirem tempo na expansão de suas fronteiras habituais (CORRALL, 2012; KIM, 2016).

O desenvolvimento das habilidades e conhecimentos relacionados à área de dados pode ocorrer em dois níveis. O primeiro, através da formação dos profissionais da informação como especialistas capazes de realizar a gestão e a curadoria de dados de pesquisa mediante educação formal em instituições de ensino e eventos da área. O segundo, através do treinamento de pesquisadores e estudantes realizado por profissionais da informação com competências e conhecimentos sobre o contexto dos dados. Neste caso, sendo realizado como parte das funções vitais das bibliotecas e serviços de informação científica de promover capacidades que permitam aos indivíduos acessarem, interpretarem, avaliarem criticamente, gerenciarem, manipularem e usarem dados de pesquisa eticamente, no qual vem sendo definido na literatura como competência em dados (CALZADA PRADO; MARZAL, 2013; KOLTAY, 2015).

Os profissionais da informação para serem capazes de ensinar a competência em dados ou estarem envolvidos em qualquer atividade de pesquisa relacionada a dados devem ser capacitados, pois o sucesso na realização dessas atividades dependerá de uma formação de qualidade. As Escolas de Ciência da Informação e os professores têm um papel importante a desempenhar no planejamento e preparação de profissionais adequadamente habilitados para atender às demandas latentes da e-Science e exercer seu papel na gestão e na curadoria de dados de pesquisa. Contudo, torna-se necessário realizar uma mudança de paradigma, pois muitas Escolas de Ciência da Informação ainda estão mais preocupadas com a informação do que com os dados. Outra maneira de melhorar a consciência dos profissionais da informação que buscam atuar na área de dados é através do estimulo dos gestores para que sua equipe realize leituras de publicações e participe de eventos sobre o tema de modo a ajudá-los a atualizar seus conhecimentos e ganhar confiança sobre o assunto (SWAN; BROWN, 2008; CORRALL, 2012; KOLTAY, 2015).

O treinamento de estudantes e pesquisadores pode ser realizado por bibliotecas e serviços de informação através de diversas atividades como a disponibilidade de orientações sobre questões relacionadas à gestão de dados, acesso aberto, comunicação acadêmica, direitos autorais, propriedade intelectual, metadados, padrões técnicos, acesso, arquivamento, reuso e preservação de dados; o desenvolvimento de serviços de curadoria digital em conjunto com centros de TI fornecendo serviços de computação, armazenamento e segurança de dados; a conscientização e a defesa em torno da gestão de dados para as comunidades de pesquisa, cujo nível de compreensão provavelmente variará no campus devido as diferentes culturas disciplinares que muitas vezes tornam os desafios da gestão de dados específicos à disciplina. Porém, a maioria das bibliotecas universitárias e científicas ainda

oferecem apenas programas de competência informacional para seus usuários, sendo incomum ver estas participando de projetos de pesquisa para auxiliarem na gestão de dados (SWAN; BROWN, 2008; CORRALL, 2012).

Diante dessas mudanças, a CI está sendo desafiada a desenvolver uma nova vertente de especialidades e funções capazes de responder às crescentes necessidades de serviços relacionados aos dados de pesquisa, além de auxiliar seus usuários a serem capazes de gerenciar os seus próprios conjuntos de dados. Para isso, torna-se necessário fornecer formação e treinamento que ajustem às novas formas de manipulação, gestão e uso dos dados de pesquisa, preparando tanto os profissionais da informação quanto os pesquisadores e estudantes de graduação e pós-graduação para lidar com o que surgirá daqui em diante, pois estamos num ambiente informacional completamente novo do que tínhamos há 10 anos (KOLTAY, 2015; KIM, 2016; TANG; SAE-LIM, 2017).

2 OBJETIVOS

O objetivo geral dessa pesquisa é identificar se existem na literatura nacional da CI estudos que contemplem a formação dos profissionais da informação em gestão e curadoria de dados de pesquisa, e estudos que tratem sobre o desenvolvimento da competência em dados para pesquisadores e estudantes de graduação e pós-graduação. Os objetivos específicos são: a) identificar trabalhos pertinentes ou relacionados às temáticas; b) avaliar como as temáticas estão sendo apresentadas nos artigos; c) caracterizar o estado atual da literatura da CI sobre as temáticas; d) promover a discussão sobre os temas na CI brasileira.

3 PROCEDIMENTOS METODOLÓGICOS

O trabalho consiste numa pesquisa qualitativa e exploratória. A partir da revisão de literatura apresentaram-se algumas contribuições relevantes para a construção teórica e conceitual da pesquisa, e através da análise de conteúdo avaliaram-se os documentos selecionados. Na primeira etapa realizou-se a coleta de dados, através da busca na literatura brasileira da CI sobre as temáticas supracitadas, em bases nacionais, BRAPCI e SCIELO, bases internacionais, LISA, LISTA, *Scopus* e *Web of Science*, além do Google Acadêmico. Para isso, usaram-se as palavras-chave "*Curriculum*" OR "Currículo", "*Information Science*" OR "Ciência da Informação", "*Curation*" OR "Curadoria" e "*Data*" OR "Dados". Na estratégia de busca almejou-se aumentar a revocação e ampliar a possibilidade de recuperação de documentos úteis.

Na segunda etapa selecionaram-se os documentos recuperados através da leitura do título, resumo, palavras-chave e texto completo para reconhecer documentos que tratam das temáticas supracitadas e/ou identificar artigos que as mencionam de algum modo, pois estas ainda são incipientes na literatura brasileira da CI. Ao final dessa etapa, permanecem 10 documentos, nos quais 07 são artigos de periódico, 01 é livro, 01 é tese e 01 é monografia.

Na terceira etapa, usou-se como critério de exclusão, analisar somente os 07 artigos de periódicos devido à sua natureza de comunicação. Em seguida, elaboraram-se as categorias para

avaliar os assuntos tratados nos artigos, indicando dessa forma a proximidade ou o distanciamento dos artigos com os objetivos almejados para a pesquisa (ver Tabela 1 - Apêndice A).

4 RESULTADOS

Os artigos são analisados nos resultados por critério de ordem cronológica.

4.1 Boeres e Cunha (2012)

Os autores citam a escassez de trabalhos que tratam sobre as competências necessárias para formação dos profissionais da informação na área de preservação digital. Apresentam as demandas para bibliotecários com domínio sobre as tecnologias digitais, gestão da informação e promoção do acesso à informação em formato impresso ou digital. Discutem sobre o desenvolvimento das capacidades analíticas do profissional que lida com os dados para tomar decisões nos negócios. Alertam quanto à necessidade de profissionais com know-how para lidar com a segurança de dados digitais. Porém, apesar de refletirem sobre algumas problemáticas em relação aos dados, os autores dedicam à formação do profissional da informação com foco a aprendizagem de técnicas de digitalização de recursos informacionais e preservação digital, e não a atuação direcionada à área de dados de pesquisa.

4.2 Curty e Serafim (2016)

Os autores estudam currículos de cursos de ciência de dados dos EUA e a formação de profissionais com competências para lidar com volumes cada vez maiores de dados em diferentes contextos. Os termos “gerenciamento de dados” e “curadoria de dados” aparecem no trabalho e indicam que a formação dos cientistas de dados também passa por essas temáticas, práticas e teorias. Ainda que não trate diretamente os dados gerados pela ciência, aborda sobre a demanda existente para os profissionais de dados neste contexto devido ao avanço da e-Science. Também apontam para a possibilidade dos cientistas de dados trabalharem com pesquisadores ou grupos de cientistas em ambientes de pesquisa e serviços que tratem de dados digitais. Entretanto, não foi objetivo da pesquisa discutir o desenvolvimento de competências em dados de pesquisa e/ou a formação dos profissionais da informação orientada aos dados de pesquisa.

4.3 Boeres e Cunha (2016)

Os autores apresentam conceitos sobre preservação digital, biblioteca digital e curadoria digital e os relacionam com as etapas de gestão de dados de pesquisa. Tratam sobre questões de

autenticidade, confiabilidade e integridade dos dados digitais. Destacam o reuso ético dos dados de pesquisa por pesquisadores como usuários de repositórios de dados e que têm por intuito retirar desses conjuntos de dados novas análises e percepções. Concluem que Escolas de Ciência da Informação são responsáveis pela capacitação dos profissionais da informação para preservação e curadoria digital, porém não foi o objetivo principal dos autores tratar sobre a formação dos profissionais da informação para o contexto dos dados de pesquisa.

4.4 Weber (2017)

O autor cita a curadoria digital como responsável pelo gerenciamento do objeto digital. Aborda sobre os repositórios de dados, o ciclo de vida dos dados, a e-Science e a importância de realizar a gestão dos dados de pesquisa. Discorre sobre a necessidade dos pesquisadores aplicarem as práticas de curadoria no seu dia-a-dia utilizando os metadados para garantir o reuso e o acesso aos dados. Discute os riscos dos repositórios digitais se tornarem cemitérios de dados, caso a prática de curadoria recaia exclusivamente sobre o sistema ao invés de envolver também a prática humana. Aponta que a curadoria digital surge principalmente como necessidade da pesquisa científica na era digital e, portanto, deve contar com profissionais capacitados. O objetivo principal da pesquisa foi discutir os conceitos de curadoria digital, e desse modo, não trata diretamente das temáticas aqui pesquisadas.

4.5 Araújo e Finamor (2017)

Os autores discutem as funções do arquivista na curadoria e na preservação digital sob a ótica da Arquivologia. Tratam da importância de se assegurar que os futuros usuários sejam adequadamente capacitados para garantir um efetivo gerenciamento e reuso dos dados, o que significa um alerta para o desenvolvimento de competência em dados por usuários desses serviços, ainda que esse não seja o foco principal do artigo. Apresentam as competências que devem ser desenvolvidas pelos profissionais interessados em atuar na curadoria e na preservação digital. Concluem sobre a necessidade de disponibilizar disciplinas de graduação nos cursos de Arquivologia do Brasil que formem profissionais aptos a trabalharem na área da curadoria e preservação digital, além de incitar profissionais pela formação continuada. Portanto, percebe-se que as temáticas aqui pesquisadas foram brevemente citadas, mesmo não sendo o objetivo do trabalho.

4.6 Rocha, Sales e Sayão (2017)

O artigo foca no uso das ferramentas de escrita e registro de procedimentos e resultados de pesquisa científica pelos pesquisadores, avaliando os seus impactos. Discute sobre essas ferramentas sendo anteriormente encontradas apenas em formato analógico, porém, atualmente existem também em formato digital, denominadas de cadernos eletrônicos de laboratório. Trata o seu uso e

as práticas relacionadas à ciência aberta e a preservação de dados institucionais, além da importância da atuação do profissional da informação nos ambientes de pesquisa como mediadores do uso dessas novas e promissoras tecnologias. Há também uma acanhada menção ao bibliotecário de dados e a formação desses profissionais que ainda está muito distante da realidade dos cursos de graduação de Biblioteconomia no Brasil.

4.7 Vanz, Santin e Pavão (2018)

O artigo tem como objetivo principal a bibliometria como serviço de informação a ser ofertado por bibliotecas universitárias brasileiras a partir de novas demandas dos usuários. Indicam que a bibliometria ainda não faz parte dos conteúdos obrigatórios da maior parte dos cursos de Biblioteconomia no país, muito menos é comum enquanto serviço das bibliotecas universitárias. Apontam brevemente sobre a necessidade dos bibliotecários auxiliarem os pesquisadores a lidarem com seus dados de pesquisa e a atenderem a demanda das agências de financiamento. Expõem sobre a importância de promover a competência informacional nas bibliotecas. Porém, não foi discutida a formação da competência em dados para pesquisadores e estudantes como parte das novas atividades dos bibliotecários das universidades.

5 CONSIDERAÇÕES FINAIS

Observam-se, atualmente, dificuldades para encontrar artigos na CI no Brasil quanto à formação do profissional da informação em gestão e curadoria de dados de pesquisa e sobre o desenvolvimento da competência de dados em estudantes e pesquisadores pelos serviços de informação científica. Não foi constatado nenhum artigo. Destaca-se que os artigos selecionados para análise tratam de modo incipiente sobre esses assuntos, pois não eram os seus objetivos principais. Apenas um artigo discutiu brevemente sobre algumas atividades relacionadas ao desenvolvimento de competência em dados de pesquisa e quatro artigos indicam sobre necessidade de formação do profissional da informação para atividades relacionadas ao contexto de dados. Considera-se a necessidade de realizar discussões e pesquisas sobre essas temáticas na literatura nacional da área de modo a tentar estimular a mudança dos currículos atuais das Escolas de Ciência da Informação e dos serviços prestados aos estudantes e pesquisadores no Brasil, pois torna-se imprescindível para a CI acompanhar as transformações na ciência e evoluir conjuntamente com as práticas de produção e comunicação científica.

REFERÊNCIAS

ARAÚJO, R. O.; FINAMOR, M. S. Curadoria digital: papéis e responsabilidades do arquivista. **Inf. Prof.,** Londrina, v. 6, n. 1, p. 44 – 68, jan./jun. 2017.

BOERES, S. A. A.; CUNHA, M. B. Competências básicas para os gestores de preservação digital. **Ci. Inf.**, Brasília, DF, v. 41 n. 1, p.103-113, jan./abr., 2012.

BOERES, S. A. A.; CUNHA, M. B. Competências para a preservação e curadoria digitais. **Rev. Digit. Bibliotecon. Cienc. Inf**. Campinas, SP v.14 n.3 p. 426-449, set/dez. 2016.

CALZADA PRADO, J.; MARZAL, M. A. Incorporating data literacy into information literacy programs: core competencies and contents. **Libri**; v. 63, n. 2, p. 123−134. 2013.

CORRALL, S. Roles and responsibilities: libraries, librarians and data. In: PRYOR, Graham (Ed.). **Managing research data**. London: Facet, 2012. p. 141-151.

CURTY, R. G.; SERAFIM, J. da S. A formação em ciência dos dados: uma análise preliminar do panorama estadunidense. **Informação&Informação**, Londrina, v. 21, n. 2, p. 307–328, maio/ago., 2016.

KIM, J. Who is teaching data: meeting the demand for data professionals. **Journal of Education for Library and Information Science Online**, v. 57, n. 2, p. 161-173, 2016.

KOLTAY, T. Data literacy: in search of a name and identity. **Journal of Documentation**, v. 71, n. 2, p. 401-415, 2015.

ROCHA, L. L.; SALES, L. F.; SAYÃO, L. F. Uso de cadernos eletrônicos de laboratório para as práticas de ciência aberta e preservação de dados de pesquisa. **PontodeAcesso**, Salvador, v.11, n.3, p. 2-16, dez. 2017.

SWAN, A.; BROWN, S. **The skills, role and career structure of data scientists and curators**: an assessment of current practice and future needs. Truro, UK: Key Perspectives, 2008.

TANG, R.; SAE-LIM, W. Data Science Programs in U.S. higher education: an interview with the authors, **Journal of eScience Librarianship**, v. 6, n. 1, p. 1-17, 2017.

VANZ, S. A. S.; SANTIN, D. M.; PAVÃO, C. M. G. A bibliometria e as novas atribuições profissionais em bibliotecas universitárias. **InCID: R. Ci. Inf. e Doc.**, Ribeirão Preto, v. 9, n. 1, p. 4-24, mar./ago. 2018.

WEBER, C. Curadoria digital de dados científicos: pelo viés de um periódico. **P2P & Inov**. Rio de Janeiro, RJ, v. 3 n. 1, set./mar. 2017.

Apêndice A

Quadro 1 – Avaliação dos artigos

Categoria	**Artigo 1** Boeres e Cunha (2012)	**Artigo 2** Curty e Serafim (2016)	**Artigo 3** Boeres e Cunha (2016)	**Artigo 4** Weber (2017)	**Artigo 5** Araújo e Finamor (2017)	**Artigo 6** Rocha, Sales e Sayão (2017)	**Artigo 7** Vanz, Santin e Pavão (2018)
Literatura da CI	Sim	Sim	Sim	Sim	Sim	Sim	Sim
Temas principais	Preservação e biblioteca digital	Ciência de dados	Preservação e curadoria digital	Curadoria de dados científicos	Curadoria e preservação de documentos digitais	Ciência aberta e preservação de dados de pesquisa	Bibliometria em biblioteca universitária
Objetivo geral	Identificar as competências dos profissionais de preservação digital	Compreender os aspectos formativos do cientista de dados	Identificar competências e atividades da preservação e curadoria digital	Estudar os conceitos de curadoria digital e os assuntos relacionados à temática	Investigar as funções do arquivista na curadoria e preservação digital	Relacionar o uso de cadernos eletrônicos com o reuso e a preservação de dados	Discutir o uso da bibliometria como atividades das bibliotecas acadêmicas
Menção a dados de pesquisa	Não	Sim	Sim	Sim	Sim	Sim	Sim
Competência em dados de pesquisa	Não	Não	Não	Não	Não	Não	Trata brevemente do assunto
Formação do PI em Gestão e Curadoria de Dados de Pesquisa	Não	Trata brevemente do assunto	Trata brevemente do assunto	Não	Trata brevemente do assunto	Trata brevemente do assunto	Não

Fonte: Dados de pesquisa, 2018.

A CIÊNCIA DE DADOS À LUZ DA CIÊNCIA DO SENSO COMUM ARISTOTÉLICO:
uma análise comparativa com a tecnologia sensorial

DATA SCIENCE BASED ON ARISTOTELIAN COMMON SENSE SCIENCE:
a comparative analysis with the robotics technology

Tereza Cristina Fernandes[1], Marcello Mundim Rodrigues[2]
(1) Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, terezacristinafernandes@gmail.com, bolsista FAPEMIG
(2) Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, marcellomundim@ufu.br

Resumo:
Compara a percepção da compreensão aristotélica aplicáveis à Ciência de Dados, com objetivo de buscar uma fundamentação teórica filosófica a esta disciplina emergente. Apoia-se em duas metodologias: (a) a revisão bibliográfica - nas contribuições dos estudos críticos pertinentes à filosofia aristotélica da experimentação do senso comum e; (b) na análise de conteúdo – pela capacidade de inferência e dedução da compreensão. Argumenta-se a partir da análise dos três níveis de pensamento observado por Aristóteles, sob os seguintes aspectos: (1) os sentidos e as sensações são os dados que a mente recebe; (2) as ideias são os dados que a mente produz e; (3) o pensamento produz conhecimento relacionando as ideias, juntando-as, separando-as e confrontando-as, formado pela capacidade de cognição em busca da razão. Como resultado comparativo, encontra similaridades com as tecnologias atuais, que captam dados do ambiente, utilizando instrumentos que imitam capacidades sensoriais do raciocínio humano, bem como a compreensão desses dados. Finalmente, considera a necessidade de identificar as raízes dos pressupostos teóricos filosóficos, a fim de consolidar a Ciência de Dados e despertar para a produção de novos conhecimentos.
Palavras-chave: Ciência de dados. Filosofia. Tecnologia sensorial.

Abstract:
The study compares the perception of the aristotelian comprehension applicable to Data Science, with objective of seeking theoretical and philosophical grounds to this emergent subject. The research bases itself on two methods: (a) literature review – on the contributions of the critical studies relevant to the aristotelian philosophy of the common sense experimentation and; (b) content analysis – for its inference capability and comprehension deduction. It argues from the three levels of thinking observed by Aristotle, under the following aspects: (1) senses and feelings are data that the mind receives; (2) ideas are data that the mind produces and; (3) thinking produces knowledge that relate ideas, combining, splitting and confronting them, in which is created by the cognition capability in search of reason. As a comparative result, it finds similarities with the current technologies, which captures data from the environment, using instruments that imitate human reasoning sensory abilities, as well as data comprehension. At last, it considers the need for identifying the roots of the theoretical and philosophical assumptions, in order to consolidate Data Science and awaken for the production of new knowledge.
Keywords: Data science. Philosophy. Sensory technology.

1 INTRODUÇÃO

A Ciência de Dados é uma área de estudo que cresceu exponencialmente nos últimos 10 anos. De maneira global, escolas de Ciência da Informação e as I-Schools tentam absorver essa disciplina, oferecendo formação à nível de graduação e pós-graduação com o intuito de profissionalizar o "cientista de dados", indicando que este conhecimento se tornou relevante ao cenário atual e uma oportunidade de crescimento profissional.

Sendo interdisciplinar, apresenta várias visões diferentes. Porém, segundo Wang (2018), há um certo consenso entre os pesquisadores no que diz respeito a identificar e extrair padrões valiosos de grandes volumes de dados, convertendo dados em informação e conhecimento através da mineração e análise de dados. O mesmo pesquisador avalia que a missão e a natureza da Ciência de Dados são consistentes com as da Ciência da Informação e que estão intimamente inter-relacionadas e, juntas, formam os componentes da pesquisa da "cadeia de informações".

Portanto, a Ciência de Dados é uma área emergente e necessária, que absorve conhecimento de outras áreas e se torna o mais novo interesse da Ciência da Informação. Todavia, é possível considerá-la uma ciência?

Do ponto de vista epistemológico, por ser uma nova disciplina e por conter peculiaridades interdisciplinares, é uma área que requer abordagens críticas que lhe dê sustentação sólida, enquanto campo científico. Embora não seja intenção deste artigo estender-se na base conceitual, os termos "dado", "datum" ou "data", os dois últimos, singular e plural em latim, possuem atributos multifacetados, caracterizando-os como um conceito ambíguo. Portanto, sendo os dados um fenômeno atual, que surge com desafios e paradoxos, torna sua ciência um estudo ainda mais reflexivo, com necessidade de sustentações filosóficas.

O pensamento aristotélico, embora muitas vezes criticado por apresentar uma rigidez ordenada e hierarquizada da natureza, começou onde todos deveriam começar – naquilo que já sabia, na sua experiência mental simples de suas observações. O saber prático necessário se baseava no conhecimento do senso comum da natureza, com o intuito de entender suas leis ocultas. Considerando o senso comum, este artigo busca doutrinas aplicáveis à Ciência de Dados, que interessam mais à história da ciência que à história da filosofia.

O quarto paradigma da ciência difundido por James Nicholas Gray (Jim Gray), há mais de uma década constrói a ciência atual a partir da junção das tradicionais que se usam da observação e experimentação (empirismo), teste de teorias, computação e suas simulações com o acréscimo de conceitos, técnicas e ferramentas analíticas dirigidas a dados.

Esse paradigma se baseia no pragmatismo científico, mas anseia por bases filosóficas sólidas para melhor entendimento desse fenômeno. A velocidade do fazer de hoje não facilita a construção de uma fundamentação desse novo paradigma, pois esse processo necessita de cognição, lógica e indagações que consomem tempo, um aspecto extremamente frágil nessa nova era pragmática dos dados.

Neste sentido, este artigo busca se sustentar na ideia de que a filosofia pode tanto argumentar como responder lacunas da Ciência de Dados. Do ponto de vista do pensamento simples, argumentações filosóficas de meras especulações podem tomar grandes proporções epistemológicas, conduzindo a experimentos práticos por cientistas.

Neste caso considera-se que a ciência precisa da filosofia tanto quanto a filosofia precisa da ciência. Em seu livro *Um discurso sobre ciências*, Boaventura de Souza Santos (2002, p. 10) propõe

um novo senso comum, em que a "[...] distinção hierárquica entre conhecimento científico e conhecimento vulgar tenderá a desaparecer e a prática será o fazer e o dizer será a filosofia prática", o que caracteriza uma certa tolerância científica em relação às experiências humanas.

A história da ciência está repleta de confirmações de como especulações filosóficas aguçaram experimentos científicos. Cita-se o exemplo da física do movimento: a falta de precisão e lógica dos argumentos da física aristotélica sobre sua observação do movimento, conduziram sucessores pesquisadores como Galileu, Descartes, Newton e Einstein ao intenso processo da evolução desse conhecimento.

Dessa forma, este artigo parte da premissa que uma ciência se fundamenta não apenas pela metodologia, mas também pela história e filosofia dos fenômenos.

2 OBJETIVOS

O presente artigo possui característica introdutória, resumindo-se a algumas observações gerais e simples, com o objetivo de motivar estudos e reflexões ulteriores.

Sem assumir posição se a ciência deve se orientar em relação ao senso comum, é objetivo deste artigo apresentar similaridades do pensamento aristotélico no que confere suas experiências do senso comum ao produzir conhecimento. Em outro plano, indicar contribuições do pensamento aristotélico sobre a ciência dos dados, contribuindo para a percepção da tecnologia sensorial. Já sob o ponto de vista da percepção do senso comum, indicar semelhanças envolvidas nos processos perceptíveis, nos processos de encadeamento lógico e na cognição através de posições teóricas sólidas fundamentadas na ciência aristotélica.

3 METODOLOGIA

Este artigo apoia-se em duas metodologias: (a) a revisão bibliográfica - nas contribuições dos estudos críticos pertinentes à filosofia aristotélica da experimentação do senso comum e; (b) na análise de conteúdo – pela capacidade de inferência e dedução da compreensão.

Metodologicamente, para a escolha dos documentos, utilizou-se a regra dos textos representativos. Após a leitura flutuante, algumas hipóteses foram levantadas. Em seguida, analisou-se o conteúdo dos livros e artigos com base nos relacionamentos entre três categorias semânticas principais: ciência, filosofia aristotélica e ciência dos dados. Para alcançar os resultados, utilizou-se a técnica de inferência que levaram às interpretações. Em seguida, foi feita uma análise comparativa para apresentar similaridades entre o pensamento aristotélico no que confere suas experiências do senso comum que buscam produzir conhecimento e suas inter-relações.

4 RESSONÂNCIAS DA CIÊNCIA NO SENSO COMUM ARISTOTÉLICO

De maneira resumida, sob o argumento da necessidade da filosofia na ciência, esta seção visa dar atenção à evolução da Ciência da Computação e mostrar como o conhecimento aristotélico, em parte, auxiliou o desenvolvimento de novas teorias, e por conseguinte, técnicas e tecnologias sofisticadas.

Francis Bacon e John Locke, filósofos do iluminismo, repudiaram completamente o pensamento aristotélico e medieval, como muitos empiristas desde então, muitas vezes negligenciaram a importância crítica da lógica aristotélica para o pensamento científico. Na época moderna, o saber prático necessário tem se baseado no conhecimento das ciências naturais, proporcionando tecnologia através do conhecimento científico (ADLER, 2010).

Porém, chama atenção a história da Ciência da Computação e suas derivações advindas da lógica formal de Aristóteles, com o raciocínio silogístico, mas interpretado e adaptado para a álgebra, no trabalho de George Boole (1815–1864). "As interpretações respectivas dos símbolos 0 e 1 no sistema de lógica são Nada e Universo", disse o matemático (GLEICK, 2013).

Dessa forma, substitui-se termos do silogismo por variáveis, e as palavras lógicas "tudo", "é" e "portanto" com operadores aritméticos. Assim, na famosa premissa maior de Aristóteles "Todos os homens são mortais", pode-se interpretar "Tudo no conjunto *x* também está no conjunto *y*", exemplificado por Boole, originando as portas lógicas mais básicas, na linguagem original de Boole, E (AND), OU (OR) e NÃO (NOT) (FERNEDA, 2012).

Incentivado por Vannevar Bush (1890–1974), a visão de Claude Shannon (1916–2001) era que o sistema de Boole poderia ser mapeado diretamente em circuitos elétricos, ao perceber que a teoria correta seria "exatamente análoga ao cálculo das proposições usadas no estudo simbólico da lógica". Ele mostrou a correspondência entre circuitos elétricos e operações booleanas e como mapear a lógica no mundo físico ao distinguir suas camadas. Por volta de 1940, Alan Turing (1912–1954) assimilou esse conhecimento e o utilizou ao projetar computadores na linguagem da lógica matemática (TASINAFO, 2008).

Portanto, pode-se inferir que a gênesis da Ciência da Computação é uma ressonância do aguçamento iniciado por Aristóteles, motivado por conceitos que nascem no cotidiano (senso comum). Seguindo seus passos, Boole propôs que as *Leis do Pensamento* começam com uma descrição de seus objetivos, para investigar as leis fundamentais da operação da mente humana, assim como proposto por Aristóteles.

A lógica concebida por Aristóteles há 25 séculos e que perdura nos dias atuais é ensinada nas aulas de filosofia e Ciência da Computação – lógica de primeira ordem ou de predicados – com algumas modificações do sistema de Gottlob Frege (1848–1925) é inspirada pela tentativa de Boole de melhorar a lógica de Aristóteles.

5 POR QUE A CIÊNCIA DOS DADOS?

Trata-se de um tema basal para a Ciência da Informação (CI). Zins (2011) define: "dados são frequentemente concebidos como matéria-prima para a informação, que é frequentemente concebida como matéria-prima para o conhecimento". Este mesmo autor afirma: "dados, informação e conhecimento são inter-relacionados" (ZINS, 2007, p. 452).

O estudo científico dos dados não é exclusivo da Ciência da Informação, pois incorpora elementos variados e se baseia em técnicas e teorias oriundas de muitos campos, como Ciência da Computação, Inteligência Artificial, engenharias, estatísticas, entre outras, sendo assim ligada em muitos campos, uma área altamente interdisciplinar. Porém, devido a essa característica, os dados, ou

o que constitui dados, são falíveis, portanto, são conjecturas. Consequentemente o Big Data é falível e contaminado por teorias (FRICKÉ, 2015).

Os dados são um novo fenômeno da contemporaneidade que proporcionam grandes desafios e oportunidades: os atuais volumes de dados a analisar, bem como seu dinamismo e capilaridade, surgem como oportunidades de progresso científico e inovação tecnológica em diferentes áreas do eixo ciência-indústria-governo.

Os dados tornam reais as possibilidades de inferência da informação. A partir do tratamento e análise dos dados ou conjuntos de dados, pode-se manipulá-los para inferir (deduzir de maneira lógica) conhecimentos que podem variar desde atividades do emissor de dados, como também conter informações sobre o próprio emissor dos dados.

No domínio universal, "dados" são um conjunto de códigos, representando percepções empíricas ou "matéria-prima empírica". No domínio abstrato ou subjetivo, "dados" são estímulo sensorial ou seu significado (ou seja, a percepção empírica) (ZINS, 2011).

Portanto, a motivação inicial da Ciência de Dados é seu próprio fenômeno na contemporaneidade, sua aplicabilidade e sua transversalidade com outras áreas, sendo assim, tudo que possa contribuir para seu estudo deve ser investigado.

Inspirado por novas teorias acerca dos dados e sua interdisciplinaridade, o artigo busca refletir – a partir de análises, inferidas por pressupostos aristotélicos – os dados sensíveis que consistem na matéria-prima do pensamento e conduzem à possibilidade de inferência, uma potência na construção do pensamento cognitivo em ato, como será visto a seguir.

6 O PENSAMENTO, O CONHECIMENTO E OS DADOS

Ao dedicar estudos sobre os homens, no célebre *Sobre a alma*, Aristóteles introduz entre outras compreensões, a descrição dos sentidos externos e de sua distinção dos sentidos internos, donde advém parte deste estudo comparativo. Outros tratados, entre eles, *Categorias*, *Da interpretação*, *Sobre a sensação,* também contêm doutrinas que interessam a este estudo, ao relacionar propriedades que são inerentes ao ser humano; a capacidade do pensamento e acúmulo de conhecimento.

As seções a seguir (6.1 e 6.2) são inferências e interpretações dos referidos tratados sobre a mente. Nessa abordagem, na visão aristotélica, três níveis correlacionados do pensamento são necessários para produzir conhecimento. O primeiro nível refere-se aos sentidos, o segundo, às ideias e o terceiro, ao pensamento.

A seção 7 corresponde aos resultados da analogia dos três níveis de pensamento com os dados e a relação com a tecnologia sensorial.

6.1 Os Dados Sensíveis, Matéria-Prima das Ideias

Os sentidos estão presentes nos órgãos receptores, respondendo a estímulos sensoriais do mundo externo. Essa é uma capacidade passiva do corpo humano, em que cada um dos sentidos

produz sensações quando o órgão sensível sofre ação física de algo do mundo exterior. As sensações podem ser vividas tanto separadamente quanto em conjunto a outros sentidos. Assim, é possível em um único objeto obter a sensação de paladar, tato e olfato ao mesmo tempo, ou produzir sensações simultâneas - sinestesia.

Entretanto, os sentidos sozinhos não são capazes de produzir ideias, eles representam as matérias-primas de como a experiência sensível é formada. Segundo Aristóteles, a experiência sensível torna-se ativa quando a percepção envolve os processos da ação da memória e da imaginação, aliadas à capacidade de compreensão das sensações do mundo exterior. Para o filósofo, o esforço da mente de compreender as informações dos sentidos e da percepção, ou seja, da experiência sensível, forma as ideias.

A partir da atividade racional, não dos sentidos, o cérebro, órgão que recebeu as instruções dos órgãos dos sentidos, consegue distinguir as capacidades motoras, do paladar, da visão, audição e tato, tais como: andar, pular, repouso, cores, sons, entre outros. A capacidade de compreender instantaneamente a natureza das diferenças entre cães e gatos, por exemplo, se dá pela ação da natureza particular de cada um deles. É na mente que a forma das coisas se tornam ideias claras, separando-as da matéria individualizante.

Aristóteles diferencia a capacidade da mente de produzir coisas da capacidade de produzir ideias. Uma é o exato oposto da outra, pois ao produzir coisas, a mente transforma a matéria de acordo com as ideias, enquanto que ao produzir ideias a mente toma as formas das coisas e as transformam em ideias.

Quando a mente forma ideias de cães e gatos, por exemplo, a compreensão desses animais está presente na experiência sensível. Pela ação da sensação, é possível perceber, lembrar, recordar, imaginar ou até sonhar com esses tipos de "objetos".

6.2 Produzir Conhecimento – O Pensamento por Cognição

Este nível de pensamento é chamado por Aristóteles de raciocínio ou de inferência, atribuído a nossa capacidade de cognição. Segundo Da Silva (2011), todo modelo de categorização é cognitivo, dado que este é um processo intrínseco ao pensar e ao agir humano. Somente quando uma asserção se torna a base para afirmar ou negar outra asserção é que a mente ascende ao terceiro nível de pensamento. Nesse nível, o pensamento envolve dar razões para aquilo que pensamos. Entretanto, também nesse nível, aquilo que pensamos pode ser lógico ou ilógico (ADLER, 2010).

O terceiro nível de pensamento está associado à teoria do silogismo, objeto de referência científica à lógica booleana, conforme visto na seção 4. Porém, para compreender o silogismo, deve-se de antemão conhecer a divisão das coisas em categorias através do método lógico dedutivo.

O discurso – técnica altamente desenvolvida, fundamentada na teoria do silogismo – deve compreender o encadeamento hierárquico por gradação, sendo a substância, o centro unificador na ontologia aristotélica. Por conseguinte, examinar quais são os pressupostos embutidos nas ideias, testar o grau de confiabilidade de cada uma e hierarquizar os conceitos na sua ordem de dependência.

7 RESULTADOS

Inter-Relação Entre o Pensamento do Senso Comum Aristotélico, a Ciência dos Dados e a Tecnologia Sensorial

Em analogia, conforme visto na seção 6.1, os sentidos e as sensações são os dados que a mente recebe, enquanto que as ideias são os dados que a mente produz.

Foram considerados dados, matérias primas com instruções captadas pelos órgãos dos sentidos que se conectam à mente. Chegando ao seu destino, as matérias primas (ou instruções) são liberadas formando ideias. Portanto, o primeiro e segundo nível de pensamento formam as ideias, as quais são os dados que a mente produz como resultado daquilo que recebe a partir dos órgãos dos sentidos.

Por vezes chamadas de "instruções", os dados também são considerados, apoiando-se em Zins (seção 5), a "matéria-prima" para a construção de ideias.

O terceiro nível se dá quando a mente começa a produzir ideias a partir da experiência sensível. Este nível depende da capacidade cognitiva de organizar as ideias através do encadeamento de objetos e palavras para poder expressá-las.

Nesta abordagem do senso comum aristotélico, surge uma questão: uma máquina com IA poderia ser programada pela percepção sensorial, utilizando scanners de alta resolução – tal qual os órgãos receptores – para entender e interpretar o implícito "cognitivamente", os dados que os sensores captam e transformá-los em uma informação concreta?

Esta percepção tecnológica já é uma realidade contemporânea. Robôs conseguem aprender baseado no que acontece no ambiente, como ocorre nos carros autômatos, por exemplo.

O plano categórico hierárquico, apresentado na seção 6.2, envolve a organização e a definição da estrutura hierárquica do universal e suas relações com os atributos. Os enormes volumes de dados podem ser lidos sequencialmente pelo computador, o qual pode efetuar operações lógicas ou algébricas em grande velocidade, unindo não só a inteligência artificial, mas também as ciências cognitivas e a Ciência da Informação. Portanto, os processadores podem relacionar as classificações (análise preditiva), análise de agrupamentos de dados tal como o encadeamento sistemático do conhecimento de Aristóteles.

Sendo assim, processadores auxiliam a organizar dados por classificação, a partir de uma análise preditiva, ou por análise de agrupamento hierárquico, por partição ou por densidade, práticas da mineração de dados. O mesmo acontece no silogismo e na lógica, que propõem o encadeamento, ou agrupar nas práticas de sistematização, a partir da observação de um discurso, em busca do aumento do grau de certeza.

8 CONSIDERAÇÕES FINAIS

Como foi visto na seção 6, Aristóteles sempre fundamentou sua ciência a partir de estudos históricos articulados com sua filosofia. Na seção 4 foi visto o exemplo de cientistas que fundamentaram

seus estudos a partir de conhecimentos aristotélicos. Na introdução, mostramos como erros da física aristotélica motivaram estudos subjacentes em busca da realidade. Isso demonstra a preocupação da ciência não só em identificar possíveis raízes históricas, como também buscar uma solidez tanto para refutar estudos anteriores como para confirmar suas hipóteses.

Todavia, um dos objetivos deste artigo foi apresentar argumentos que identifiquem raízes filosóficas históricas da Ciência de Dados para que essa se sustente, aproximando-a de ferramentas tecnológicas que utilizam os dados como sua principal matéria-prima.

A Ciência de Dados, enquanto campo científico e do conhecimento, também necessita de uma solidez na sua fundamentação. Outro objetivo deste artigo foi trazer uma sustentação filosófica à nova área de conhecimento que vem emergindo, tal qual orienta o quarto paradigma difundido por Jim Gray.

Outro objetivo deste artigo foi comparar os estímulos externos que os órgãos receptores recebem a dados, indicando semelhanças aos dados codificados de computadores com tecnologia sensorial. De maneira similar, os dados sensoriais são instruções da percepção enviada ao cérebro, enquanto que os dados da tecnologia sensorial são dados enviados à processadores tecnológicos (seção 6.1).

No contexto tecnológico foram observadas semelhanças no que concerne os procedimentos utilizados pela análise de dados utilizados na (IA) com os princípios da realidade categórica, cognitiva e lógica metafísica de Aristóteles (seção 6.2).

A abordagem filosófica se sustentou nos tratados de Aristóteles, por este ter sido o pioneiro "cientista" a descrever, usando experiências mentais simples, o comportamento sensorial orgânico humano.

Muitas vezes, a capacidade sensitiva é tão natural aos seres humanos que não prestamos atenção em como esse fenômeno ocorre. Hoje, encontramos similaridades com a ciência moderna pelo complexo sistema nervoso central. Células nervosas levam as "instruções", ou as "matérias primas", ou "códigos", ou "células", em última instância, os "dados", que irão se encaixar por sinapses nas terminações nervosas mentais. Consequentemente o sistema nervoso decodifica esses dados que recebe do meio externo, que podem ser arquivados ou respondidos se solicitados. Podem ser simples reações como movimentos e sensações agradáveis, como também formulação de conhecimento complexo.

Nesta perspectiva, há um corolário de comparações que podem ser feitas sob o ponto de vista do papel dos dados nos processos de diversos paradigmas da ciência. Neste artigo, foi abordado o da tecnologia sensitiva, que copia, de maneira efetiva, quase biológica, os processos atribuídos aos órgãos sensíveis humanos.

Dessa forma, como foi verificado, existem certas correspondências da percepção sensitiva da observação do senso comum, que, quando comparadas criticamente, podem ser transformadas em tecnologias empíricas. Assim, também, através de percepções da Organização do Conhecimento, da representação das categorias cognitivas como forma de encadear e identificar os dados.

Contudo, ainda estamos longe de um robô que entenda completamente o mundo. Sonhos, esperanças, sentimentos, em escala menor, é o que nos distingue.

Embora fazer comparações de dados – a partir do ponto de vista da tecnologia sensorial – possa parecer coerente, este não é um ponto de partida que auxilie no avanço científico para a

Ciência de Dados. Inicialmente, deve-se verificar o fenômeno em proporções tanto históricas quanto futuristas.

Por fim, deve-se mencionar que a ciência moderna tenta descrever as condições materiais subjacentes a certos efeitos empíricos, enquanto a ciência aristotélica procura explicar as coisas em termos de propósitos e os tipos de coisas que existem. De fato, visam coisas diferentes, porém, isso não significa que um refuta o outro, eles se sustentam na busca do conhecimento.

REFERÊNCIAS

ADLER, M. J. **Aristóteles para todos:** uma introdução simples a um pensamento complexo. São Paulo: É Realizações, 2010.

ARISTÓTELES. **Poética**: Organon; Política; Constituição de Atenas. São Paulo: Nova Cultural, 1999. 315 p. (Os pensadores).

_____ . **Categories**. Da alma (De anima), int., trad. e notas por Carlos Humberto Gomes. **Lisboa: Edições**, v. 70, 2001.

_____. **Metafísica**: Livro I. Trad. e nota de Lucas Angioni. Campinas: IFCH/UNICAMP, 2002.

COTRIM. G. **Fundamentos da filosofia**: história e grandes temas. 15. ed. São Paulo: Saraiva, 2002.

DA SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. Brasil: Elsevier, 2017.

DA SILVA, A. R.; D. O. LIMA, G. A. B. As categorias e a categorização cognitiva contemporânea: enfoque sobre os olhares da biblioteconomia e ciência da informação e das ciências cognitivas. **Revista Interamericana de Bibliotecología**, v. 34, n. 3, 2012.

FERNEDA, E. **Introdução aos modelos computacionais de recuperação de informação**. Ciência moderna, 2012.

FRICKÉ, M. Big data and its epistemology. **Journal of the Association for Information Science and Technology**, v. 66, n. 4, p. 651-661, 2015.

GLEICK, J. **A informação: uma história, uma teoria, uma enxurrada**. Editora Companhia das Letras, 2013.

HEY, T.; TANSLEY, S.; TOLLE, K. (Editors). **The fourth paradigm**: data-intensive scientific discovery. Redmond, Washington: Microsoft Research, 2009.

REALE, G. **Introdução a Aristóteles**. Rio de Janeiro: Contraponto, 2012.

SANTOS, B. de S. **A crítica da razão indolente**: contra o desperdício da experiência. 3. ed. São Paulo: Cortez, 2001.

TASINAFO, P. M. Um breve histórico do desenvolvimento da lógica matemática e o surgimento da teoria da computação. **Anais do 14° Encontro de Iniciação Científica e Pós-Graduação do ITA – XIV ENCITA / 2008**. São Paulo, 2008.

WANG, L. Twinning data science with information science in schools of library and information science. **Journal of Documentation**, v. 74, n. 6, p. 1243-1257, 2018.

ZINS, C. Conceptions of information science. **Journal of the American Society for Information Science and Technology**, v. 58, n. 3, p. 335-350, 2007.

_____. Redefinindo a Ciência da Informação: da "ciência da informação" para a "ciência do conhecimento". **Informação & Sociedade: Estudos**, v. 21, n. 3, 2011.

A CONSTRUÇÃO DO REPOSITÓRIO DE DADOS DA UFPB:
a experiência com o dataset de Arboviroses

THE CONSTRUCTION OF THE UFPB DATA REPOSITORY:
the experience with the Arboviruses Dataset

Pollianna Marys de Souza e Silva[1]
Sandra de Albuquerque Siebra[(2)]
(1) UFPB. pollianna_marys@hotmail.com
(2) UFPE. sandra.siebra@gmail.com

Resumo:
Repositório de Dados (RD) são sistemas digitais de informação que armazenam dados provenientes de pesquisas científicas, a fim de colaborar com o seu acesso, uso, reuso e preservação. Nesse contexto, este artigo objetiva relatar a experiência de construção do RD da UFPB. Esta foi uma pesquisa-ação, qualitativa e descritiva. Como resultados iniciais se obteve a instalação da plataforma Dataverse e criação do RD da UFPB, no ano de 2018, tendo como conjunto de dados inicial um dataset construído a partir da extração de mais de um milhão de posts da rede social Twitter, de outubro de 2017 a março de 2018, sobre as arboviroses, que é um conjunto de patologias formado principalmente pela Zica, Dengue e Chikungunya. Com esse dataset espera-se disponibilizar dados brutos que possam servir para pesquisadores de várias áreas estudarem a disseminação das doenças, do que se fala sobre elas, mapear seus agravos e surtos, entre outros.
Palavras-chave: Repositório de Dados. Dataverse. Dataset Arboviroses. Curadoria de Dados. Ciência Aberta.

Abstract:
Data Repository (RD) are digital information systems that store data from scientific research in order to collaborate with its access, use, reuse and preservation. In this context, this article aims to report on the experience of building the UFPB's data repository. This was an action research, qualitative and descriptive. As initial results we obtained the installation of the Dataverse platform and creation of the RD of the UFPB, in the year of 2018, having as dataset a dataset constructed from the extraction of more than one million posts of the social network Twitter, of October of 2017 to March 2018, on arboviruses, which is a set of pathologies formed mainly by Zika, Dengue and Chikungunya. With this dataset, it is expected to provide raw data that can be used for researchers from various areas to study the spread of diseases, what is being said about them, map diseases and their outbreaks, among others.
Keywords: Data Repository. Dataverse. Arboviruses Dataset. Data Curation. Open Science.

1 INTRODUÇÃO

O movimento em favor do acesso aberto, conhecido como *Open Access*, surgiu a partir da crise dos periódicos e permitiu a democratização do acesso à informação, sendo os repositórios digitais uma das primeiras plataformas digitais de acesso aberto (SILVA JÚNIOR; BORGES, 2014). Estes são ambientes digitais que possibilitam reunir dados e informações de cunho científico, administrativo, técnico, artístico, cultural, entre outros, cuja função principal é promover a visibilidade de seus objetos digitais, preservando-os por meio do gerenciamento de informação (ABREU; VIDOTTI, 2016). Eles podem ser de vários tipos (SILVA JÚNIOR; BORGES, 2014): a) Temáticos - armazenam conteúdos de assuntos especializados; b) Nacional - agrega a produção acadêmica e/ou científica de um país; c) Institucional – foca na produção intelectual de uma instituição; e d) De pesquisa ou de dados – armazenam dados de pesquisa, sendo esses últimos o foco da pesquisa que originou esse artigo.

De acordo com Sayão e Sales (2016), os repositórios de dados (RD) garantem os princípios de transparência e oferecem um sistema de armazenamento seguro, além da possibilidade de se ter os dados de pesquisa disponíveis on-line, indexados, documentados, para serem acessados, baixados, visualizados e processados por pessoas ou por sistemas, estendendo-os a uma comunidade mais ampla e conectada em rede. De fato, os RD têm sua importância como recurso informacional e se tornam um dispositivo de troca de experiências e compartilhamento de dados científicos[1], além de parte essencial das infraestruturas mundiais de pesquisa em escala global, tornando visível e aberta para toda a sociedade uma parcela importante da atividade de pesquisa, caracterizando a chamada *Open Science* ou Ciência Aberta[2] (FORMENTON, 2015).

Sayão e Sales (2016) consideram que os RD, de forma diferente das publicações acadêmicas que falam por si próprias, precisam ter explícitos os seus conteúdos, para poder revelar e transmitir conhecimento no tempo e no espaço, de forma que os dados possam ser interpretados, sintetizados e reanalisados em contextos diversos, com finalidades diferentes das quais foram gerados e coletados originalmente. Assim como é preciso garantir a gestão e preservação a longo prazo, o que faz com que seja necessário o desenvolvimento de metodologias (gerenciais e tecnológicas) que guiem desde a criação dos dados, passando por seu gerenciamento, armazenamento, análise, uso e reuso, em uma diversidade de contextos. O que vem sendo proporcionado pela curadoria de dados, que "engloba atividades de gestão requeridas para manter e gerir dados de pesquisa a longo prazo, de modo que estejam disponíveis para o reuso, favorecendo a colaboração entre pesquisadores e o avanço da ciência" (SIEBRA; BORBA; MIRANDA, 2016, p. 12-13).

Entre as plataformas disponíveis para criação de um RD encontra-se o Dataverse. Ele é um RD de código aberto para compartilhar, preservar, citar, explorar e analisar dados de pesquisa. Pesquisadores, autores de dados, editores, distribuidores de dados e instituições afiliadas recebem o crédito apropriado por meio de uma citação de dados com um identificador persistente (por exemplo, DOI ou *Handle*) (RICE; SOUTHAL, 2016). Cada conjunto de dados contém metadados descritivos e arquivos de dados (incluindo documentação e código que acompanham os dados) (DATAVERSE PROJECT, 2018). Essa foi a plataforma utilizada nessa pesquisa.

1 Dados científicos são os materiais comumente registrados e aceitos na comunidade científica como necessários para validar os resultados de pesquisa. Ex: fatos e estatísticas recolhidas para posterior referência ou análise, documentos, questionários, transcrições, algoritmos, scripts, etc. (DUDZIAK, 2016).

2 Movimento para tornar a pesquisa científica, os dados e a divulgação destes acessíveis a todos os níveis da sociedade (WIKIPÉDIA, 2018).

2 OBJETIVOS

Esse artigo teve por objetivo relatar a experiência de criação do 1º RD da UFPB, fazendo uso da plataforma Dataverse, com foco no *dataset* de arboviroses. Esse RD surgiu após a realização do projeto de pesquisa - Chamada Universal - MCTI/CNPq (Número 01/2016) – intitulado "A Ciência da Informação e a Disseminação de Informações Associadas à Epidemia de Zika Vírus: uma investigação baseada na Análise de Redes Sociais". E a motivação para o projeto surgiu após o Brasil vivenciar, em 2016, um surto de microcefalia em bebês recém-nascidos, causada pela contaminação de mulheres grávidas com o vírus zyca. Sendo que para a criação do *dataset* decidiu-se expandir para um conjunto de patologias formado principalmente pela Zyca, Dengue e Chikungunya, denominadas arboviroses. Isso porque essas são doenças com graves repercussões para a saúde da população conforme indicadores operacionais e epidemiológicos.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa que originou esse artigo trata-se de uma pesquisa-ação, qualitativa e descritiva (MICHEL, 2009). Na pesquisa-ação os pesquisadores-participantes estão envolvidos de modo cooperativo e participativo na pesquisa, que é concebida e realizada em estreita associação com uma ação ou resolução de um problema coletivo (BALDISSERA, 2001). Na pesquisa qualitativa, segundo Michel (2009), existe a ligação e correlação de dados interpessoais, coparticipação das situações dos informantes e estes são analisados a partir da significação que eles dão aos seus atos. E é descritiva na medida que colhe, interpreta e discute fatos e situações. Além disso, essa é uma pesquisa bibliográfica que fez uso de artigos, livros, teses e dissertações, assim como documental pois fez-se necessário estudar o *site* e o manual da plataforma Dataverse.

O *dataset* foco desse relato é um conjunto de dados composto por posts da rede social *Twitter* sobre as arboviroses, que engloba a Zyca, Dengue e Chikungunya. Assim, a coleta de dados para compor o dataset foi realizada no período de outubro de 2017 a março de 2018. Os dados foram coletados por um script feito na linguagem de programação Ruby, fazendo uso de API (*Application Programming Interface*) disponibilizada pelo próprio *Twitter*. Os *posts* coletados foram os que apresentaram uma ou mais das seguintes palavras, desconsiderando maiúsculas e minúsculas: zica, zika, zyca, zkv, zikav, dengue, dengue hemorrágica, chikungunya, chicungunya, arbovirose, arboviros, microcefalia.

Os *posts* identificados, mais de um milhão, foram baixados no formato JasonB e categorizados considerando 3 grupos: Zyca, Dengue e Chikungunya. Em seguida, foi criado um banco de dados modelado para as normas do Dataverse, onde os dados extraídos foram efetivamente armazenados. Ressalta-se que, para poder contextualizar e adicionar valor aos dados, foi criado um conjunto de metadados descritivos para sintetizar os diferentes contextos em que as mensagens foram utilizadas e compreender como os usuários da rede social discutiram/escreveram sobre a temática. O Dataverse foi escolhido como plataforma por ser uma das mais popularmente utilizadas para criação de RDs.

Para realização do trabalho, a equipe foi composta por 6 pessoas: 1 coordenador e 5 bolsistas das áreas de Ciência da Informação e Ciência da Computação.

4 RESULTADOS

O RD da UFPB foi disponibilizado em 2018, no endereço https://dataverse.ufpb.br/dataverse/root. A equipe inicial de pesquisadores em uma primeira etapa precisou se familiarizar com o que é e como funciona um repositório de dados, o que foi sanado com a pesquisa bibliográfica. Posteriormente, sobre a plataforma Dataverse. Como toda documentação da plataforma estava em inglês, a equipe se dedicou a traduzir e destacar as principais partes da documentação, antes de começar a implantação da plataforma na instituição. Uma vez a plataforma instalada, ela foi populada com o *dataset* de arboviroses extraído do *Twitter*.

Para Gomes (2013, p. 29): "[...] a otimização e atualização de dispositivos e ações de informação têm de responder às demandas intensivas de informação e comunicação que satisfaçam, em quantidade e qualidade, às metas e finalidades da promoção da saúde e do atendimento clínico". O que ressalta as possibilidades de colaboração entre a Ciência da Informação e a área de Saúde. O que pode ser concretizado por exemplo, com a exploração, coleta e organização de bases de dados que possam colaborar com pesquisas na área de saúde, como é o caso do *dataset* de arboviroses.

Para um melhor gerenciamento de acesso e segurança, buscou-se restringir o acesso aos dados a pesquisadores previamente cadastrados, cujas credenciais de acesso podem ser feitas a partir do cadastro individual do pesquisador no Dataverse, utilizando uma conta do Gmail. Assim, optou-se por permitir acessar o arquivo restrito do *dataset* apenas após login e senha com as credenciais de acesso adequadas ao sistema e, durante *download* do arquivo do dataset, o sistema deve solicitar a aceitação de um termo de acesso e uso. O objetivo é manter um controle de acesso aos dados disponibilizados. Adicionalmente, para aqueles que navegarem nos datasets sem baixar nenhum arquivo há a solicitação de registro no *Guestbooks* (livro de visitas) de cada dataset. Ressalta-se que o acesso aos dados no RD Dataverse ocorre de quatro formas: aberto, fechado, gerenciado e moderado, o que pode trazer flexibilidade para os casos de dados de pesquisa que necessitem ficarem em sigilo por algum motivo (pesquisa em andamento, patenteamento de solução, etc).

O repositório criado a partir da plataforma Dataverse possui a identificação especificada no Quadro 1.

Quadro 1 – Identificação do Dataverse UFPB.

Dataverse	Departamento de CI
Identifier	https://dataverse.ufpb.br/dataverse/dci
Category	Department
Affiliation	CCSA - UFPB
Description	Este repositório concentra dados de pesquisas do Departamento de Ciências da Informação (DCI) do Centro de Ciências Sociais Aplicadas (CCSA) da Universidade Federal da Paraíba (UFPB).

O campo "Dataverse" descreve o título do banco de dados, o departamento, a universidade ou a revista que conterá os dados. O "Identifier" refere-se a um nome curto a ser utilizado como URL (*Uniform Resource Locator* - Localizador Padrão de Recursos que é um formato de atribuição universal para designar um recurso na Internet). A "category" identifica o tipo de grupo ao qual a base de dados pertence, no caso, departamento. O campo "Affiliation" deve compor algum nome ou sigla associado à instituição proprietária do repositório, no caso CCSA (Centro de Ciências Sociais Aplicadas) da Universidade Federal da Paraíba (UFPB) e, por fim, "Description" que deve ser a descrição resumida da base de dados, especificando o que é armazenado nela, para exibição na página do repositório.

No Dataverse, os campos de metadados são escolhidos para uso em cada conjunto de dados (dataset) a serem adicionados (THE DATAVERSE PROJECT, 2018). Considerando a natureza multidisciplinar das pesquisas do Departamento de CI, os pesquisadores optaram pelo uso de campos de metadados gerais, em um padrão que pudesse ser aplicado em dados das diversas áreas do conhecimento. Assim, os principais campos a serem preenchidos no Dataverse para o *dataset* são: título, autor(es), informações para contato, descrição resumida da pesquisa, data da coleta dos dados no RD, tópicos da pesquisa, palavras chave, entre outros. Esses metadados tanto contextualizam e descrevem os dados armazenados, como facilitam a sua recuperação.

A preservação e acesso a longo prazo são garantidos no Dataverse pela identificação persistente, que protege os documentos digitais com mecanismos que preveem a obsolescência dos dados - migração dos dados para um *software* mais recente e a prescrição que consiste em guardar o conjunto de *bytes* para serem consultados quando for necessário. (DATAVERSE PROJECT, 2018).

4 CONSIDERAÇÕES FINAIS

Nesse cenário, as arboviroses são enfermidades tropicais endêmicas, que merecem receber atenção especial dos profissionais de saúde que atuam na atenção básica e na vigilância em saúde e dos gestores das esferas federal, estadual e municipal (FIOCRUZ, 2017). Elas incapacitam ou matam milhões de pessoas e representam uma necessidade médica importante que permanece não atendida. Assim, o estudo sobre a disseminação de informações sobre essas enfermidades, tanto nos círculos acadêmicos e científicos formais, como nos informais, na mídia de massa através das redes sociais, apresenta inúmeras oportunidades de investigação para pesquisadores. É nesta vertente que a proposta do *dataset* arboviroses da UFPB está direcionada, contribuindo com dados brutos para o campo científico e para a sociedade diante da urgência da temática. Podendo-se investigar: o que a população sabe sobre as arboviroses? Que tipo de dúvidas possuem? Como a informação tem chegado até a sociedade? Que tipo de queixas são realizadas sobre as doenças em questão? Que localidades mais discutem sobre arboviroses? Além de ser possível mapear casos de relatos de morte, surtos e agravamento das doenças por meio dos posts coletados.

Ressalta-se que essa pesquisa ainda está no início, porém, a implantação e configuração do dataverse e a coleta dos dados do *Twitter* com a criação do *dataset* de arboviroses já se mostram como primeiros resultados, que podem trazer diversos desdobramentos no futuro.

Espera-se, como trabalhos futuros, poder relatar os usos feitos do *dataset* de arboviroses tanto pelos administradores do Dataverse, como por pesquisadores cadastrados na plataforma.

REFERÊNCIAS

ABREU, J. P.; VIDOTTI, S. A. B. G. Curadoria Digital nos Contexto dos Repositórios Digitais. In: Encontro Internacional de Dados, Tecnologia e Informação, 2., 2016. Marília. **Anais...** Marília: UNESP, 2016.

FIOCRUZ. **Agência Fiocruz de Notícias**. Disponível em: <https://agencia.fiocruz.br/doen%C3%A7as-negligenciadas>. Acesso em: 08 jun. 2018.

BALDISSERA, A. Pesquisa-Ação: uma metodologia do conhecer e do "agir" coletivo. **Sociedade em Debate**, Pelotas, v. 7, n. 2, p. 5-25, agosto, 2001. Disponível em: <http://revistas.ucpel.edu.br/index.php/rsd/article/viewFile/570/510>. Acesso em: 02 ago. 2018.

DATAVERSE PROJECT. **About Dataverse**. Disponível em: <https://dataverse.org/>. Acesso em: 08 jun. 2018.

FORMENTON, D. **Identificação de Padrões de Metadados para Preservação Digital**. São Carlos: UFSCar, 2015. 102 f. Dissertação de Mestrado - Universidade Federal de São Carlos. Disponível em: <https://repositorio.ufscar.br/bitstream/handle/ufscar/7221/DissDF.pdf?sequence=1>. Acesso em: 02 jun. 2018.

GOMEZ, M. N. G. O Domínio das Informações em Saúde. In: PINTO, V. B.; CAMPOS, H. H. (Org). **Diálogos Paradigmáticos Sobre Informação para a Área da Saúde.** Fortaleza: Edições UFC, 2013.

MICHEL, M. H. **Metodologia e Pesquisa Científica em Ciências Sociais**. 2 ed. São Paulo: Atlas, 2009.

RICE, R; SOUTHALL, J. **The Data Librarian's Handbook.** London: Facet, 2016. 169p.

SAYÃO, L. F.; SALES, L. F. Algumas Considerações Sobre os Repositórios de Dados de Pesquisa. **Informação & Informação**, Londrina, v. 21, n. 2, p. 90 – 115, maio/agosto, 2016. Disponível em: <http:www.uel.br/revistas/informacao/90>. Acesso em: 12 jul. 2018.

SIEBRA, S. A.; BORBA, V. R.; MIRANDA, M. J. K. F. O. Curadoria digital: um termo interdisciplinar. In: XVII Encontro Nacional de Pesquisa Em Ciência da Informação (ENANCIB), 17., 2016, Salvador. **Anais...** Salvador, BA: UFBA, 2016. Disponível em: <http://www.ufpb.br/evento/lti/ocs/index.php/enancib2016/enancib2016/paper/view/4107/2559>. Acesso em: 2 jul. 2018.

SILVA JUNIOR, L. P.; BORGES, M. M. Preservação digital no Repositório Científico de Acesso Aberto de Portugal. **Rev Eletrônica de Comun. Inf. Inov. Saúde**, v. 8, n. 4, p. 567-574, out./dez. 2014. Disponível em:<http://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/911>. Acesso em: 04 ago. 2018.

WIKIPÉDIA. **Open Science.** Disponível em: <https://en.wikipedia.org/wiki/Open_science>. Acesso em: 12 ago. 2018.

A DISSEMINAÇÃO DE INFORMAÇÕES ASSOCIADAS A EPIDEMIA DO ZIKA VÍRUS:
Uma pesquisa a partir de dados coletados no *Twitter*

THE DISSEMINATION OF INFORMATION ASSOCIATED WITH THE ZIKA VIRUS EPIDEMIC:
A research based on data collected from the Twitter

Guilherme Ataíde Dias[1], Denysson Axel Ribeiro Mota[2], André Luiz Dias de França[3], Eduardo de Santana Medeiros Alexandre[4]
(1) Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Paraíba, guilhermeataide@ccsa.ufpb.br.
(2) Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Paraíba, denysson.mota@ufca.edu.br.
(3) Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Paraíba, andreluizjpb@gmail.com.
(4) Instituto Federal da Paraíba, eduardo.ufpb@gmail.com.

Resumo:
A pesquisa apresentada, no momento em andamento, investiga de que maneira a Ciência da Informação pode fornecer suporte para entendimento de questões que emergem a partir da disseminação de informações acerca do Zika vírus postadas na rede social *Twitter.* O objetivo principal é investigar a disseminação de informações através de canais formais e informais como redutora de incertezas de uma população imersa em epidemia de Zika Vírus. A pesquisa configura-se com relação aos procedimentos técnicos como bibliográfica, com relação aos objetivos é exploratória e possui natureza quantitativa e qualitativa. O *corpus* estudado constituiu-se de mensagens disseminadas (*tweet*) através da rede social *Twitter* com conteúdos relacionados ao Zika Vírus no período compreendido entre outubro de 2017 até março de 2018. A partir dos dados coletados é possível mencionar que os usuários do *Twitter* que postam mensagens com o termo "zica" e *strings* equivalentes, não estão na maioria das vezes fazendo referência às questões de saúde associadas com a epidemia de Zika Vírus. O termo zica (zika) é usado em uma grande variedade de contextos que transcendem o domínio específico da área da saúde. A continuidade da pesquisa possibilitará uma compreensão mais aprofundada do fenômeno.
Palavras-chave: Twitter; Disseminação de dados; Zika vírus.

Abstract:
The ongoing research presented investigates how Information Science can provide support for understanding issues that emerge from the dissemination of information about the Zika virus posted on the social network Twitter. The main objective is to investigate the dissemination of information through formal and informal channels as an element capable of reducing of uncertainties on a population immersed in a Zika Virus epidemic. The research is characterized in relation to the technical procedures as being bibliographic, with regard to the objectives is exploratory and has a quantitative and qualitative nature. The studied corpus consisted of tweet messages disseminated through the social network Twitter with contents related to the Zika Virus in the period between October 2017 until March 2018. From the data collected it is possible to mention that Twitter users who post messages with the term "zica" and equivalent strings are not often referring to the health issues associated with the Zika Virus epidemic. The term zica (zika) is used in a wide variety of contexts that transcend the specific domain of health. The continuity of the research will allow a deeper understanding of the phenomenon.
Keywords: Twitter; Data dissemination; Zika virus.

1 INTRODUÇÃO

Vários temas relacionados com a área da saúde possuem ligação com a Ciência da Informação. Exemplos de trabalhos que referendam este pensamento podem ser encontrados nos materiais organizados por Duarte e Farias (2009), Galvão e Ricarte (2012) e Pinto e Campos (2013).

De forma mais ampla ilustramos a essência das intersecções entre a Ciência da Informação e a área da saúde através do discurso de Gomez (DIAS, 2016):

> A convergência de questões de informação e questões em saúde é de longa data, sendo a área de informação em saúde rica e pioneira na inovação e pesquisa em informação e em tecnologias de informação. A otimização e atualização de dispositivos e ações de informação têm de responder às demandas intensivas de informação e comunicação que satisfação, em quantidade e qualidade, às metas e finalidades da promoção da saúde e do atendimento clínico. A Ciência da Informação compõe uma das muitas redes inter e transdisciplinares onde essas questões e temáticas são objeto de pesquisa e reflexão (GOMEZ, 2013, p.29).

É de amplo conhecimento que o mundo, e de forma mais marcante os países da América Central e do Sul estão sendo atingidos por uma série de enfermidades atribuídas ao Zika Vírus (Vide Figura 1), como por exemplo, a microcefalia e a *Síndrome de Guillain–Barré (DIAS, 2016)*. Este é um exemplo de fenômeno que pode ser estudado através do prisma da Ciência da Informação.

Figura 1: Distribuição do Zika Vírus - Américas

Fonte: Organização Mundial de Saúde Apud Dias (2016)

O Zika Vírus tem como vetores de transmissão os mosquitos denominados de *Aedes aegypti* e *Aedes albopictus,* o primeiro com abrangência apenas em regiões tropicais e subtropicais, o segundo podendo sobreviver e hibernar em regiões de clima mais frio (WHO, 2016).

No Brasil**,** a eclosão de casos de doenças hipoteticamente associadas ao Zika Vírus, de forma mais notória a microcefalia, tem como epicentro os estados de Pernambuco e da Paraíba (DIAS, 2016).

A disseminação de informações sobre o Zika Vírus, tanto nos círculos acadêmicos e científicos formais, como nos informais na mídia de massa através das plataformas de rede social, apresentam inúmeras oportunidades de investigação para o pesquisador em Ciência da Informação. É nesta vertente que esta pesquisa está estruturada, através da análise de dados através da Análise de Redes

Sociais (ARS) e de ferramentas de visualização que possibilitam a compreensão da disseminação de informações relacionadas com a epidemia de Zika Vírus, contribuindo com os resultados obtidos para o campo científico e para a sociedade, visto a urgência e relevância da temática (DIAS, 2016).

2 OBJETIVOS

A pesquisa em tela investiga de que maneira a Ciência da Informação pode fornecer suporte para entendimento de questões que emergem a partir da disseminação de informações acerca do Zika vírus postadas na rede social *Twitter*. O objetivo principal é investigar a disseminação de informações através de canais formais e informais como redutora de incertezas de uma população imersa em epidemia de Zika Vírus. Os objetivos específicos são: Mapear o fluxo informacional sobre o Zika Vírus no âmbito do *Twitter*; Identificar as especificidades da disseminação das informações acerca do Zika Vírus no *Twitter*; Desenvolver *software* para coletar dados sobre o tema Zika Vírus no *Twitter*; Produzir visualizações de dados a partir dos dados coletados para gerar novos padrões associativos; Gerar métricas associadas aos dados coletados no *Twitter* (DIAS, 2016).

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa ora em andamento, configura-se com relação aos procedimentos técnicos como bibliográfica, com relação aos objetivos é exploratória e possui natureza quantitativa e qualitativa (GIL, 2008).

O *corpus* estudado constituiu-se de mensagens disseminadas (*tweets*) através da rede social *Twitter* com conteúdo relacionado ao Zika Vírus. Diversos padrões de grafias associados ao Zika Vírus foram utilizados na captura das mensagens. Esta estratégia teve como por objetivo possibilitar a coleta do maior número de mensagens possível relacionadas com o Zika Vírus.

Para a captura dos *tweets* foi elaborado um *script* na linguagem de programação *Ruby*. Esta linguagem foi escolhida por permitir o desenvolvimento rápido de aplicações através de acesso simplificado a *Application Programming Interface* (API) do *Twitter* e por ser de fácil aprendizado. Os *tweets* foram obtidos no padrão *JavaScript Object Notation* (JSON).

O *script* para a captura dos *tweets* foi implantado em um servidor *Debian Linux* e capturou mensagens postadas no período compreendido entre outubro de 2017 até março de 2018, integralizando um total de 6 (seis) meses.

Os *tweets* capturados foram armazenados em banco de dados *PostgreSQL* através do tipo de dados JSON.

O *script* capturou durante o seu período de execução um total de 2.907.720 *tweets*, das quais mais de 82% são mensagens escritas nas línguas portuguesa, inglesa e espanhola. As consultas ao *corpus* foram feitas através da *Structured Query Language* (SQL).

Indicamos como obstáculo para a captura dos *tweets* as eventuais quedas do servidor *Debian Linux* em virtude de interrupções no fornecimento de energia elétrica na Universidade Federal da Paraíba nos finais de semanas e as interrupções no acesso à internet. Esta situação foi minimizada com ajustes

na configuração de parâmetros entre o sistema operacional *Debian Linux* e a aplicação *Ruby* para que este segundo tentasse se recuperar após alguma condição de instabilidade do sistema.

Os *tweets* capturados foram processados mediante o uso da Linguagem SQL, no ambiente provido pelo Sistema Gerenciador de Bancos de Dados (SGBD) *PostgreSQL*, com a função de exportação dos resultados em formato CSV. Os arquivos com os dados eram então, posteriormente, visualizados ou tratados no *SublimeText* e/ou *OpenOffice*.

De modo a facilitar a análise dos *tweets*, foram criados dois arquivos em formato texto. Um arquivo constituído de *tweets* na língua portuguesa e outro composto por *tweets* na língua inglesa. Devido à estrutura JSON do *tweet* conter, em caso de *retweet*, toda a estrutura de dados do *tweet* original, o comando SQL buscava por combinações distintas de data/hora e texto, conforme exposto no código da Figura 2. Com isto, garante-se que os *tweets* originais apareçam apenas uma vez, mas também permite que no caso de dois *tweets* onde o texto seja o mesmo, por ter os dados de data/hora distintos, ambos sejam recuperados.

Figura 2: Comando SQL para recuperar data/hora de criação e texto do *tweet*

```
select
  distinct(
  dados->'created_at', texto)
from tweets
where dados->>'lang' = 'en' or
      dados->>'lang' = 'pt');
```

Fonte: Desenvolvimento nosso

Para recuperar as *hashtags* de cada *tweet*, foi utilizado o comando SQL aqui exibido na Figura 3.

Figura 3: Comando SQL para recuperar *hashtags* dos *tweets*

```
select dados->'lang',
  upper(jsonb_array_elements(
  dados->'entities'->'hashtags')->>'text'),
  count(*)
from tweets
where
 dados->'entities'->'hashtags' != '"" and
 dados->'entities'->'hashtags' != '"[]""
group by 1, 2;
```

Fonte: Desenvolvimento nosso

4 RESULTADOS PRELIMINARES

As análises iniciais dos arquivos de *tweets* revelaram alguns fatos curiosos. As postagens de *tweets* na língua portuguesa continham pouca relação com a questão da Zika e assuntos que remetesse ao tema saúde, a vasta maioria das postagens estão relacionadas com o uso do termo "Zika" em questões coloquiais do dia-a-dia das pessoas. O Quadro 1 ilustra esta situação.

Quadro 1: Amostra dos conteúdos das mensagens dos *tweets* capturados

Mensagem do *tweet*
Hoje a sessão foi zika
@saladeredacao Caiu por causa do Baldasso, homem Zika
Quando eu casar vou entrar na igreja com um Nike Zika no pé, nada de sapato social...
essa semana que vou ficar sozinho em casa vai ser mais que zika
será q saiu minha zika do fifa
festinha ficou zika ta, curti
meu gatinho curioso vai dar dengue de tão parado
Hoje foi zika
Meu treinador é Deus, que me botou pra jogar olhou pro campo e disse zica vai lá
Minha vida sexual tá mais parada que Água pra dengue
Hoje a sessão foi zika

Fonte: Dados de pesquisa

O arquivo texto com *tweets* em língua portuguesa também foi investigado para se identificar quais as *tags* mais frequentes no contexto das postagens associadas ao tema Zika. O Quadro 2, apresentado a seguir ilustra os resultados encontrados.

Quadro 2: *Hashtags* mais frequentes nos *tweets* em Língua Portuguesa

Hashtag	Qtd
MALHAÇÃO	3023
FLAMENGO	1860
DENGUE	1840
ZIKA	1787
ZICO	1770
ZICAÇO	1759
ZICÃO	1751
BBB18	1666
VIVAZICO	1612
TREZEDIAS-LONGEDOSOL	863

Fonte: Dados de pesquisa

Além das informações relacionadas com o sentido das mensagens disseminadas no *Twitter* e das *hashtags* mais frequentes encontradas no *corpus*, os pesquisadores associados à pesquisa estão investigando parâmetros relacionados com a localização geográfica das mensagens e a aderência dos *tweets* a assuntos que remetam a questões de saúde.

5 CONSIDERAÇÕES PARCIAIS

Até o estágio atual do desenvolvimento da pesquisa os objetivos específicos elencados foram parcialmente atingidos. Em uma primeira instância é possível mencionar que os usuários do *Twitter* que postam mensagens com o termo "zica" e *strings* equivalentes, não estão na maioria das vezes fazendo referência às questões de saúde associadas com a epidemia de Zika Vírus. O termo zica (zika) tem uso em uma grande variedade de contextos que transcendem o domínio específico da área da saúde. Este uso mostrou-se predominante.

É preciso empreender estudos mais aprofundados para entender os motivos que levam a população de Língua Portuguesa (Brasil) a não usar o *Twitter* como uma ferramenta que possibilite a disseminação de mensagens acerca de um problema de saúde pública, ferramenta esta que poderia ser útil na obtenção de informações sobre o combate ao mosquito vetor do Zika Vírus, bem como no tratamento das enfermidades associadas a este vírus.

REFERÊNCIAS

DIAS, G. A. **A Ciência da Informação e a disseminação de informações associadas a epidemia de Zika vírus: uma investigação baseada na Análise de Redes Sociais**. João Pessoa. Universidade Federal da Paraíba. 2016. Projeto de Pesquisa. Chamada Universal – MCTI/CNPq N. 01/2016

DUARTE, Z.; FARIAS, L. (Orgs). **A medicina na era da informação**. Salvador: EDUFBA, 2009.

GALVAO, M. C. B.; RICARTE, I. L. M. **Prontuário do Paciente**. Rio de Janeiro: Guanabara Koogan, 2012.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2008.

GOMEZ, M. N. G. O Domínio das Informações em Saúde. In: PINTO, V. B.; CAMPOS, H. H. (Org). **Diálogos Paradigmáticos sobre informação para a área da saúde**. Fortaleza: Edições UFC, 2013.

PINTO, V. B.; CAMPOS, H. H. (Org). **Diálogos Paradigmáticos sobre informação para a área da saúde**. Fortaleza: Edições UFC, 2013.

WHO. (2016). **Zika virus disease: Questions and answers**. Disponível em: <http://www.who.int/features/qa/zika/en/>. Acesso em: 03 fev. 2018.

ANÁLISE DA RELAÇÃO ENTRE PERFIL E DESEMPENHO ACADÊMICO DOS ALUNOS MATRICULADOS NA DISCIPLINA DE INTRODUÇÃO À PROGRAMAÇÃO UTILIZANDO ALGORITMOS DE CLASSIFICAÇÃO

ANALYSIS OF THE RELATIONSHIP BETWEEN PROFILE AND ACADEMIC PERFORMANCE OF STUDENTS ENROLLED IN THE PROGRAMMING INTRODUCTION COURSE USING CLASSIFICATION ALGORITHMS

Roberto Silva de Oliveira Júnior, Douglas Valentim de Almeida Cardins, Wenddell Juler Pereira de Lima, Thereza Patrícia Pereira Padilha, Vanessa Farias Dantas
Centro de Ciências Aplicadas e Educação – Universidade Federal da Paraíba (UFPB)
58297-000 – Rio Tinto – PB – Brasil
{roberto.silva, douglas.valentim, wenddell.juler, thereza, vanessa}@dcx.ufpb.br

Resumo:

Ao ingressar na Universidade, a maioria dos alunos enfrenta dificuldades no aprendizado dos conteúdos da disciplina de Introdução à Programação, primeiro componente curricular de programação nos cursos de Computação e Informática. Muitos alunos terminam evadindo ou reprovando vários semestres nesta disciplina e, assim, comprometem a conclusão de seus cursos. Sabe-se, com base em estudo da literatura, que isso se deve principalmente a aspectos como falta de motivação dos alunos, deficiências do ensino médio e dificuldades com raciocínio lógico e conceitos abstratos. Entretanto, outros fatores podem interferir nos resultados dos discentes. Diante disso, o presente artigo tem como proposta utilizar algoritmos de classificação, disponíveis na ferramenta Weka, para investigar se existe relação entre o perfil (social, acadêmico e econômico) dos alunos e o seu desempenho acadêmico nessa disciplina, com o objetivo de guiar as ações da coordenação do curso e do corpo docente no combate à evasão e à reprovação. Para isso, foram testados dados coletados de um curso de Licenciatura em Ciência da Computação da Universidade Federal da Paraíba. Foram encontradas regras que classificam os alunos em aprovado ou reprovado. Assim, foi constatado que alunos com gasto elevado com transporte, sem conhecimento prévio de computação e idade maior ou igual a 19 anos são propensos a reprovação, já os que não trabalham e têm idade menor ou igual a 18 anos e têm conhecimento prévio de computação são mais propensos a aprovação na disciplina.

Palavras-chave: Programação. Evasão. Reprovação. Mineração de dados.

Abstract:

Upon entering the University, most of the students show difficulties to learn contents of the Introduction to Programming course, first programming course in the Computer Science Degree. Many students dropped off or failed several semesters in this course and, thus delayed their degrees. In the literature, several studies show that the main factors are: lack of motivation of the students, deficiencies of high school and difficulties with logical reasoning and abstract concepts. However, other factors may interfere with student outcomes. The present paper shows the use of classification algorithms, available in the Weka tool, to investigate if there is a relationship between the students (social, academic and economic) profile and their academic performance in this course, in order to guide actions of the course coordination and faculty to minimize evasion and retention. So, data were collected of a Computer Science course from Federal University of Paraiba. Rules were found that classify students as sucess or fail. So, students with high transportation costs, no prior knowledge of computing and age greater than or equal to 19 years possibly will be to fail, and students that do not work and are less than or equal to 18 years old and have previous knowledge of computation are predisposed for sucess in the course.

Keywords: Programming. Evasion. Retention. Data mining.

INTRODUÇÃO

Uma preocupação nas Instituições de Ensino Superior (IES) no Brasil e no mundo é a evasão universitária, definida como a saída do aluno de seu curso, sem concluí-lo. Sabe-se, ainda, que os índices de evasão são maiores nos primeiros períodos dos cursos (Bardagi e Hutz, 2009). No caso dos cursos da área de Computação e Engenharia, outra preocupação é o alto índice de retenção em disciplinas de programação. Nos últimos anos, pesquisadores têm concentrado estudos na busca de encontrar respostas para estes problemas analisando dados acadêmicos, sociais e econômicos (Silva, 2013; Araújo et al., 2011; Campello e Lins, 2008).

Segundo Fagundes et al. (2014), o desempenho acadêmico constitui também um potente indicador de qualidade dos cursos de uma instituição, ou seja, sinaliza se os objetivos estão sendo alcançados. Portanto, entende-se que há uma grande pertinência na realização de estudos sobre desempenho acadêmico no curso de Licenciatura em Ciência da Computação (LCC) da Universidade Federal da Paraíba, a fim de estabelecer ações para aumentar/manter a sua qualidade que, atualmente, possui conceito máximo (5) atribuído pelo Ministério da Educação.

O aumento da disponibilidade dos dados dos alunos e o interesse de pesquisadores das áreas de Mineração de Dados e Informática na Educação, principalmente, fizeram surgir uma nova área de pesquisa conhecida como Mineração de Dados Educacionais (MDE). MDE visa explorar dados do ambiente educacional para descobrir padrões (isto é, conhecimento oculto/escondido nos dados) para contribuir no entendimento do processo de aprendizagem e, assim, auxiliar as tomadas de decisão sobre novas práticas pedagógicas (Baker, 2010). Embora MDE ainda seja pouco explorada no Brasil, esta é vista como promissora na compreensão de problemas na área da Educação (Gottardo et al., 2013).

1. OBJETIVOS

Este artigo tem como objetivo geral investigar uma relação entre o perfil (social, cultural, acadêmico e econômico) dos alunos e o seu desempenho acadêmico na disciplina de Introdução à Programação, do curso de Licenciatura em Ciência da Computação, a fim de minimizar os altos índices de evasão e de reprovação.

Em termos de objetivos específicos, tem-se:

- categorizar, em um nível maior de detalhes, o perfil dos alunos ingressantes no curso de Licenciatura em Ciência da Computação da Universidade Federal da Paraíba;
- realizar experimentos com o conjunto de dados construído (perfil, econômico e acadêmico) utilizando três algoritmos de classificação;
- identificar variáveis que influenciam os alunos no sucesso e insucesso da disciplina de Introdução à Programação (retenção e evasão);
- apresentar aos gestores acadêmicos e ao professor da disciplina os resultados da pesquisa.

2. PROCEDIMENTOS METODOLÓGICOS

No meio do semestre letivo 2017.1, foi aplicado um formulário web para 42 alunos matriculados na disciplina de IP com 19 perguntas objetivas envolvendo aspectos pessoais (Q1 a Q6), econômicos (Q10 a Q12) e acadêmicos (Q13 a Q19). Uma identificação sobre o conteúdo de cada pergunta é apresentada a seguir:

- Q1 - gênero;
- Q2 - idade;
- Q3 - estado civil;
- Q4 - possui filhos;
- Q5 - possui computador;
- Q6 - possui internet em casa;
- Q7 - cidade de origem;
- Q8 - moram na cidade em que estudam;
- Q9 - gastos mensais com transportes;
- Q10 - trabalha;
- Q11 - recebe ajuda de custos;
- Q12 - horas semanais de trabalho;
- Q13 - tipo de escola frequentada no ensino médio;
- Q14 - cursou a disciplina de IP antes de entrar no curso;
- Q15 - motivo pelo qual desistiu ou reprovou anteriormente na disciplina;
- Q16 - conhecimento da língua inglesa;
- Q17 - conhecimento de computação antes de entrar no curso;
- Q18 - conhecimento do perfil do curso de LCC;
- Q19 – LCC foi a primeira opção no Enem.

Todas as perguntas do formulário fizeram parte do conjunto de dados obtido como atributos. Ao fim do semestre, foi disponibilizada a situação final dos alunos (aprovado ou reprovado), que foi adicionado ao conjunto de dados como atributo-classe (rótulo). Dos 42 registros, 34 eram da classe aprovado e 8 reprovado. Em seguida, os dados foram transformados e disponibilizados no formatado .arff, formato padrão de entrada para a ferramenta Weka.

Por fim, os dados foram aplicados aos algoritmos de classificação OneR, PART e JRip, disponíveis na ferramenta Weka, e que têm como saída um conjunto de regras.

3. RESULTADOS

Nesta seção serão apresentados os resultados obtidos utilizando os algoritmos OneR, PART e JRip.

3.1 OneR

A Figura 1 apresenta as quatro regras geradas pelo algoritmo OneR. Neste caso, somente o atributo gastoComTransporte foi usado para separar as classes aprovado e reprovado. Onde classificou trinta e cinco instâncias corretamente (83.3%) e sete incorretamente (16.6%).

Figura 1: Resultado Obtido pelo Algoritmo OneR.

```
=== Classifier model (full training set) ===

gastoComTransporte:
        Alto    -> Reprovado
        Baixo   -> Aprovado
        Medio   -> Aprovado
        Zero    -> Aprovado
(35/42 instances correct)
```

Observa-se que alunos com gasto elevado com transporte têm sido classificado como reprovado. Analisando o conjunto de dados, verifica-se que esses alunos moram em outras cidades e, até mesmo, em outros estados, e, assim, precisam se deslocar para as suas cidades diariamente. Sabe-se que este deslocamento no cotidiano acarreta em cansaço e falta de tempo para estudos e, consequentemente, influencia no desempenho das disciplinas cursadas.

3.2 PART

A Figura 2 apresenta as quatro regras geradas pelo algoritmo PART. Neste caso, o algoritmo classificou 36 instâncias corretamente (85.7%) e 6 incorretamente (14.3%).

Figura 2: Resultado Obtido pelo Algoritmo PART.

```
PART decision list
------------------

trabalha = Nao AND
idade <= 18: Aprovado (20.0)

conhecimentoPrevioDeComputacao = Sim: Aprovado (13.0/2.0)

trabalha = Nao AND
conhecimentoPrevioDoCurso = Sim: Aprovado (5.0/2.0)

: Reprovado (4.0)

Number of Rules  :      4
```

Foram determinadas três regras para classificar os alunos como aprovados. De modo geral, o perfil dos alunos aprovados são: 1) não trabalham e possuem idade menor ou igual a 18; 2) tem conhecimento prévio sobre computação; e 3) não trabalham e tem conhecimento prévio do curso que ingressaram. Caso não esteja dentro de um desses perfis, o algoritmo determina a classe como reprovado.

3.3 JRip

Na Figura 3 podem ser observadas as duas regras extraídas pelo algoritmo JRip. Este algoritmo classificou 35 instâncias corretamente (83.3%) e 7 incorretamente (16.7%).

Figura 3: Resultado Obtido pelo Algoritmo JRip.

```
JRIP rules:
===========
(conhecimentoPrevioDeComputacao = Nao) and
(idade >= 19) => Class=Reprovado (9.0/3.0)
 => Class=Aprovado (33.0/2.0)
```

Neste caso, basicamente, alunos são classificados como reprovados, aqueles que não possuem conhecimento prévio de computação e com idade maior ou igual a 19.

A partir dos resultados obtidos com a ferramenta de mineração de dados Weka, pôde-se encontrar fatores que influenciam a reprovação e a aprovação de alunos matriculados em IP. De modo geral, gasto elevado com transporte, não conhecimento prévio de computação e idade maior ou igual a 19 anos podem levar a uma reprovação. Por outro lado, não trabalhar e ter idade menor ou igual a 18 anos, ter conhecimento prévio de computação são fatores que contribuem para uma aprovação na disciplina.

4. CONSIDERAÇÕES FINAIS

O crescente número de reprovações e evasões na disciplina de Introdução à Programação é alarmante e cabível de medidas que visem cessar tal crescimento. Dessa forma, foi primordial a captação de dados, a averiguação e o estudo deles para se obter respostas sobre os casos. Assim, a aplicação de Mineração de Dados apresentou-se como alternativa para obtenção de informações pertinentes.

Embora o conjunto de dados aplicado tenha sido restrito, principalmente, no que diz respeito à quantidade de registros de alunos reprovados, foi possível ter um modelo de perfil de alunos que perdem a disciplina de IP. Neste experimento não houve a presença de nenhum registro de aluno como evadido. Apesar de ter tido casos de alunos nesta situação durante o período avaliado, não foi possível obter seus dados pessoais, econômicos e acadêmicos destes alunos porque a coleta aconteceu no meio do semestre e os alunos já tinham desistido da disciplina. Portanto, para evitar estas perdas de dados, recomenda-se coletar os dados dos alunos até a segunda semana após o início das aulas.

REFERÊNCIAS

ARAÚJO, E. A. T; CAMARGOS, M. A.; CAMARGOS, M. C. S. Desempenho Acadêmico de Discentes do Curso de Ciências Contábeis: uma Análise dos seus Fatores Determinantes em Uma IES

Privada. **XXXV Encontro da Associação Nacional de Pós-graduação e Pesquisa em Administração**. Rio de Janeiro - RJ, Brasil, 2011.

BAKER, R. S. J. Data mining for education. **International Encyclopedia of Education**, v. 7, n. 3, p. 112 – 118, 2010.

BARDAGI, M. P.; HUTZ, C. S. **Não havia outra saída": percepções de alunos evadidos sobre o abandono do curso superior**. Psico-USF (Impr.) vol.14 no.1 Itatiba Jan./Apr. 2009.

BARROSO, M. F; FALCÃO, E. B. M. **Evasão Universitária: O Caso do Instituto de Física da UFRJ**, IX Encontro Nacional de Pesquisa em Ensino de Física. Rio de Janeiro, 2004.

BRANCO NETO, W. C.; SCHUVARTZ, A. A. Ferramenta Computacional de Apoio ao Processo de Ensino-Aprendizagem dos Fundamentos de Programação de Computadores. **18º Simpósio Brasileiro de Informática na Educação**. São Paulo, Brasil. 2007.

CAMPELLO, A. V. C.; LINS, L. N. Metodologia de Análise e Tratamento da Evasão e Retenção em Cursos de Graduação de Instituições Federais de Ensino Superior. **XXVIII Encontro Nacional de Engenharia de Produção**. Rio de Janeiro, RJ, Brasil, 2008.

DIAS, A. F. M.; CERQUEIRA, G. S.; LINS, L. N. Fatores Determinantes da Retenção Estudantil em um Curso de Graduação em Engenharia de Produção. **COBENGE 2009 – XXXVII Congresso Brasileiro de Ensino de Engenharia**. Recife – PE, 2009.

FAGUNDES, C. V.; LUCE, M. B.; ESPINAR, S. R. O desempenho acadêmico como indicador de qualidade da transição Ensino Médio-Educação Superior. **Ensaio: aval. pol. públ. Educ.**, Rio de Janeiro, v.22, n. 84, p. 635-670, jul./set. 2014.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: an overview. **Advances in Knowledge Discovery and Data Mining**, AAAI Press / The MIT Press, MIT, Cambridge, Massachusetts, and London, England, p.1-34, 1996.

GOTTARDO, E.; KAESTNER, C.; NORONHA, R. V. **Aplicação de Técnicas de Mineração de Dados para Estimativa de Desempenho Acadêmico de Estudantes em um AVA Utilizando Dados com Classes Desbalanceadas**. ICBL2013 – International Conference on Interactive Computer aided Blended Learning, p. 24-28, Florianópolis-SC, Brasil, 2013.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **The weka data mining software: an update**. SIGKDD Explor. Newsl., ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov 2009. ISSN 1931-0145.

LODER, L. L., NAKAO, O. S. **Evasão e Retenção em Cursos de Engenharia**. Sessão Dirigida. XXXIX Congresso Brasileiro de Educação em Engenharia (COBENGE 2011). Blumenau – SC, 2011.

MARQUES, A. F. **Aplicação de clusterização de dados na base de dados do saneamento ecológico e econômico de Minas Gerais**. 88p. Monografia (Bacharelado em Sistemas da Informação) – Universidade Federal de Lavras, Lavras/MG, 2014.

SCHULTZ, M. R. O., **Metodologias para Ensino de Lógica de Programação de Computadores**. Monografia de Especialização (Ciência da Computação). Universidade Federal de Santa Catarina (UFSC). 69p. Florianópolis, SC, Brasil. 2003.

SILVA, R. F. **Fatores que Influenciam o Desempenho Acadêmico.** Dissertação de Mestrado (Economia). Instituto de Ensino e Pesquisa (Insper). 42p. São Paulo, SP, Brasil, 2013.

SILVA FILHO, R. L. L; MONTEJUNAS, P. R., HIPÓLITO, O.; LOBO, M. B. C. M. **A evasão no ensino superior brasileiro**. Caderno de Pesquisa, v. 37, n.132, p. 641-659. São Paulo/SP. 2007.

SOARES, I. S. **UFRJ - Escola Politécnica - Vestibular 1993-2009 - Revisão Histórica - Vagas, Evasão e Retenção**. In: COBENGE 2009 - XXXVII Congresso Brasileiro de Educação em Engenharia. Recife, PE, 2009.

ANÁLISE DE SENTIMENTOS:
Identificando sentimentos em comentários da Rede Humaniza SUS

SENTIMENT ANALYSIS:
Identfying sentiment in comments on Humaniza SUS network

Eduardo Alves Silva[1], Luis Felipe Rosa de Oliveira [(2)]
(1) Universidade Nova de Lisboa - NOVAIMS, Campolide - Lisboa, easilva91@gmail.com
(2) Universidade Federal de Goiás, Goiânia, Goiás, luisfelipeprf@gmail.com

Resumo:
A análise de sentimentos e áreas correlatas ao processamento de linguagem natural tem-se tornado cada vez mais comuns em diversos contextos, seja no empresarial ou acadêmico. Esse tipo de análise facilita a compreensão a respeito de opiniões e sentimentos de diferentes indivíduos em relação a determinado produto ou temática. Ao levar em consideração a crescente difusão das redes sociais e a interação entre usuários nessas redes, a captação de dados para análise de sentimento se tornou mais simples do que era anos atrás, onde frequentemente eram utilizados questionários ou formulários para captar a opinião de uma pessoa a respeito de algo. Com esse intuito e com a alta densidade de informação da rede humaniza SUS (RHS), a primeira rede social brasileira criada a partir de uma política pública, voltada para o tema da humanização da saúde. A rede humaniza SUS torna-se um ambiente de estudo promissor para análise de sentimentos, fazendo-se uso de uma abordagem não supervisionada que visa o uso de um dicionário ou léxico de sentimentos criado com palavras em Português, e com a possibilidade de aplicação utilizando a linguagem de programação python, para apresentar a classificação de sentimentos dos comentários feitos pelos usuários da rede.
Palavras-chave: Rede Humaniza SUS. Análise de Sentimentos. Léxico de Sentimentos.

Abstract:
The sentiment analysis and areas related to the natural language processing has become increasingly common in several contexts, be it business or academic. This type of analysis facilitates the understanding of the opinions and feelings of different individuals regarding a given product or thematic. By taking into account the growing diffusion of social networks and the interaction among users in these networks, capturing data for feeling analysis has become simpler than it was years ago, where questionnaires or forms were often used to capture a person's opinion about something. With this intention and with the high density of information of the humaniza SUS network (RHS), the first Brazilian social network created from a public policy, focused on the humanization of health. The humaniza SUS network becomes a promising study environment for analysis of feelings, making use of an unsupervised approach that aims to use a dictionary or sentiment lexicon created with words in Portuguese, and with the possibility of application using the programming language python, to present the classification of feelings of the comments made by the users of the network.
Keywords: Humaniza SUS Network. Sentiment Analysis. Sentiment Lexicon.

1 INTRODUÇÃO

A crescente popularização das redes sociais, blogs e plataformas que incentivam a participação e colaboração de usuários, tornou-se comum permitindo o compartilhamento de mensagens que refletem suas opiniões e por vezes sentimentos a respeito de determinada temática ou produto.

Nesse sentido, a análise de sentimentos, ou mineração de opinião, corresponde ao problema de tentar identificar ou extrair emoções, opiniões ou pontos de vista expressados em um texto (DUARTE, 2013).

Liu (2012) menciona que esse tipo de análise está ligado a linguística e ao processamento de linguagem natural (PLN), gerando grande impacto nessa área do conhecimento, podendo ter grandes reflexos também em estudos sobre ciências políticas, economia, e ciências sociais, uma vez que são afetados pela opinião das pessoas.

O uso da análise de sentimentos é extenso, sendo usado para verificar o quanto um produto é bem aceito, de acordo com as opiniões de usuários, ou para compreender o sentimento gerado por uma publicação no Facebook, ou uma hashtag no Twitter. Essa variabilidade torna o uso da análise de sentimentos e opinião, um amplo campo de estudo.

Para a definição de sentimentos em um texto, normalmente é utilizado um léxico de sentimento, servindo como um dicionário de polaridade (FREITAS; VIEIRA, 2015), que agrega palavras que comumente expressam sentimentos positivos ou negativos (LIU, 2012). Atualmente os léxicos e modelos de análise de sentimento na língua inglesa são bastante comuns e assertivos em relação a análise de opinião e sentimento, no entanto, em Português são conhecidos quatro léxicos: OpLexicon, SentiLex, Brazilian Portuguese Linguistic Inquiry and Word Count (LIWC) e Onto.PT (FREITAS; VIEIRA, 2015).

O presente trabalho, tem por objetivo identificar os sentimentos expressados por comentários, fazendo uso do OpLexicon, da primeira rede social do Brasil criada no âmbito de uma política pública. Trata-se da rede humaniza SUS (RHS)[1] que se encontra em atividade desde o ano de 2008, atualmente com mais de 30 mil usuários por todo o Brasil, onde muitos são responsáveis pelas mais de 14 mil publicações existentes na rede e os quase 40 mil comentários, existentes nessas publicações.

A rede humaniza SUS, traz consigo uma temática bastante especifica, a humanização da saúde. Nesse sentido, foi verificado que tipo de sentimentos/opiniões os usuários da rede expressam a partir dos comentários, servindo também como um experimento para o uso de análise de sentimentos em Português.

2 ANÁLISE DE SENTIMENTOS

A análise de sentimento é, talvez, a aplicação mais popular da análise de texto, com um grande número de tutoriais, sites e aplicativos que se concentram em analisar o sentimento de vários recursos textuais, desde pesquisas corporativas até análises de crítica de filmes (SARKAR, 2016).

Em sua utilização a análise de sentimento tem uma série de fatores a serem considerados para que se consiga resultados que possam ser considerados coerentes, passando pela fase de pré-

1 Site da RHS – www.redehumanizasus.net

processamento do texto, até o uso de abordagens para identificação de sentimentos, potencialmente utilizando métodos supervisionados que envolvem aprendizado de máquina e não supervisionados, que faz uso de bancos de dados, ontologias ou léxicos de sentimento.

Um dos principais objetivos da análise de sentimentos é analisar um determinado gênero de texto para compreender o que ele expressa. Nesse sentido alguns fatores devem ser considerados, como a polaridade do sentimento e sua subjetividade, de acordo com Sarkar (2016) a análise de sentimentos funciona melhor em textos que têm um contexto subjetivo (opiniões) do quem em textos objetivos (fatos).

Vale ressaltar que, a simples busca por palavras como "bom" ou "ruim" não é o suficiente para para expressar o sentimento em um texto (DOSCIATTI; FERREIRA, 2013).

2.1 Léxico de Sentimento

Um léxico de sentimento pode ser um dicionário, vocabulário ou conjunto de palavras que tem uma polaridade positiva ou negativa atribuída (SARKAR, 2016). Para o presente trabalho foi utilizado um método não supervisionado. Sendo assim utilizado o léxico OpLexicon.

O OpLexicon é constituído de um total de 32.191 itens (24.475 adjetivos e 6.889 verbos, Tabela 1), tendo sua construção feita com base em textos jornalísticos e resenhas de filmes escritas em Português do Brasil, além do uso de tesauros e a tradução do léxico de opinião em inglês (SOUZA et al, 2012).

Tabela 1 – Quantidade de palavras por tipo OpLexicon

Tipo	Quantidade
adj (adjetivo)	24.475
vb (verbo)	6.889
hashtag	471
vb (verbo) det (determinante) n (nome) prp (preposição)	103
vb (verbo) n (nome) prp (preposição)	91
vb (verbo) adj (adjetivo	74
emoticons	66
vb (verbo) adv (adverbio)	22

Fonte: OpLexicon v3.0

3 PROCEDIMENTOS METODOLÓGICOS

O primeiro passo para se atingir o objetivo proposto foi captar os comentários da RHS em um recorte de 9 anos, a partir do banco de dados foram extraídos comentários de 2008 até o fim de 2017, o que gerou um total de 36.346 comentários.

Após essa etapa, foi iniciado o tratamento do texto dos comentários, utilizando a linguagem de programação python, juntamente com as etapas de sumarização e normalização de texto.

Segundo Sakar (2016) a normalização de texto é um processo de limpeza, normalização e padronização de dados com técnicas de remoção de símbolos e caracteres especiais, remoção de tags HTML (*Hypertext Markup Language*), remoção de stop words (preposições, pronomes, artigos), correção de grafia, *stemming* (reduzir palavras ao seu radical) e lematização (reduzir a flexão das palavras).

Os itens de normalização utilizados neste trabalho foram:

1. Remoção de tags HTML.
2. Remoção de simbolos e caracteres especiais.
3. Remoção de palavras irrelevantes (*Stopwords*).
4. Lematização.

Uma vez que os comentários da RHS seguem um padrão de escrita compreensível e coerente, itens como a correção gráfica não se mostra de extrema importância nessa análise, por outro lado, os comentários tem inúmeras tags HTML que foram tratadas durante o processo (Apêndice A).

Por sua vez, seguindo as outras etapas da normalização, como a remoção de caracteres especiais que removem itens como (“”, ?, !, /, ;, :), e a remoção de palavras irrelevantes, que é feita a partir de um dicionário de artigos, preposições, pronomes (a, e, é, ali, uns), entre outros, com isso o texto passa a ser transformado novamente, como é demonstrado no Apêndice B.

Vale mencionar que as *stopwords*, podem seguir determinados padrões, nessa análise foi usado um padrão comum da língua portuguesa e foram adicionados alguns termos que se encontravam nos comentários e não agregavam valor para a análise final.

A lematização por sua vez, se trata do da remoção de afixos das palavras, fazendo com que essa palavra volte para sua base raiz (SARKAR, 2016), levando em consideração a parte do discurso em que a palavra se encontra, ou seja, caso a palavra se encontre na posição de verbo dentro de um texto, ela será lematizada de uma forma, caso seja um adjetivo será de outra forma e assim por diante.

Uma vez que para a lematização é necessário identificar a que parte do discurso a palavra pertence e após isso lematizar a palavra, é possível fazer o uso de diferentes bibliotecas e ferramentas do python para essa tarefa, neste trabalho foi utilizada a biblioteca *Spacy*[2], que trabalha com uma variedade de idiomas incluindo o Português.

A lematização se faz importante pois, como demonstrado o léxico utilizado tem um número relevante de palavras que são representadas como parte do discurso, como adjetivos e verbos, conseguir lematizar o texto impacta no resultado final de identificação do sentimento de cada comentário.

Para identificar os sentimentos, foi utilizada uma abordagem similar ao utilizado para verificação de sentimentos com o léxico de opinião (HU;LIU, 2004), que se trata de um léxico com palavras positivas (1) e negativas (-1), entranto, o OpLexicon traz palavras neutras (0), ao adaptar o script da biblioteca NLTK[3] que faz uso do léxico de opinião foi possível identificar o sentimento dos comentários.

2 Biblioteca Spacy - https://spacy.io/

3 NLTK (Natural Language Toolkit) é uma biblioteca em Python que reúne funções e *corpus* textuais para tratamento e análise de texto. - https://www.nltk.org/

De forma geral, o script acessa cada um dos comentários, separa as palavras, verifica se essa palavra se encontra no dicionário léxico e qual sua polaridade, após esse processo, é feita uma contagem e caso existam mais palavras positivas no comentário, o mesmo é considerado positivo e assim por diante.

4 RESULTADOS

O trabalho com texto gera resultado diversos, que vão além da identificação de sentimentos, dessa forma foi feita uma verificação nos textos e palavras que geraram uma visão descritiva dos dados. Como por exemplo as palavras mais comuns encontradas nesses mais de 36 mil comentários (Apêndice C).

Além das palavras mais comuns é possível visualizar a interação em relação ao número de palavras utilizadas nos comentários e qual seria a média padrão de palavras.

De forma geral, foram encontrados mais de 25 mil comentários que foram classificados como positivos, 4,594 negativos e em torno de 4.692 comentários considerados neutros.

Faz-se importante ressaltar que o resultado final não apresenta indicadores de um modelo de aprendizado de máquina como, accuracy, f1-score, entre outros pelo fato de todo o trabalho ter sido desenvolvido em volta de um léxico de sentimento, o que foi feito em si representa a identificação de sentimentos nos comentários de acordo com palavras que estão presentes no léxico de sentimento, não sua classificação, a partir desse trabalho o treino de um classificador de sentimentos se dá em uma etapa posterior.

4 CONSIDERAÇÕES FINAIS

O estudo de identificação de sentimentos nos comentários da rede humaniza SUS, apresenta um resultado que demonstra estar de acordo com aquilo que a rede se propõe, que seria a colaboração entre os usuários em torno de uma temática específica.

Identificar sentimentos positivos em sua grande maioria pode ser visto como um incentivo para que o trabalho de desenvolvimento da rede tenha continuidade, foi possível perceber durante o estudo que o teor dos comentários em sua maioria, são de mensagens de agradecimento, por divulgação de encontros ou iniciativas relacionadas a humanização da saúde no Brasil, gerando um assim um ambiente de aprendizado e gratidão por todos aqueles que estão envolvidos nesse processo.

Os resultados obtidos em relação a esse estudo demonstram uma abordagem inicial em relação a tudo aquilo que pode ser produzido a partir do uso da análise de sentimentos e processamento de linguagem natural. A rede humaniza sus é um ambiente próspero para a possível criação de um léxico de sentimento relacionado a saúde e a humanização da saúde.

Essa mesma abordagem gera caminhos para que, a partir dos resultados desse estudo, seja criado um classificador de sentimentos utilizando métodos supervisionados e aprendizado de máquina, o que daria um contexto mais abrangente para os resultados, bem como para o uso dos dados da RHS.

Por fim, este trabalho agrega um valor experimental para o desenvolvimento de análises de sentimento utilizando o Português brasileiro, com ênfase em um método não supervisionado e com a possibilidade de construção de dados para sua posterior aplicação em algoritmos e classificadores de aprendizagem de máquina.

REFERÊNCIAS

DOSCIATTI, M. M.; FERREIRA, E. C. L. P. C. Identificando emoções em textos em português do brasil usando máquina de vetores de suporte em solução multiclasse.In: **ENIAC-Encontro Nacional de Inteligência Artificial e Computacional.** Fortaleza, Brasil, 2013.

DUARTE, E. S. **Sentiment analysis on twitter for the portuguese language.** Tese (Doutorado) — Faculdade de Ciências e Tecnologia, 2013.

FREITAS, L. d; VIEIRA, R. Exploring resources for sentiment analysis in portuguese language. In: IEEE. In: 2015 **Brazilian Conference on Intelligent Systems**. BRACIS. [S.l.], 2015. p. 152–156.

HU, M; LIU, B. Mining and summarizing customer reviews. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, Seattle, Washington, p.168-177, 2004

LIU, B. **Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies.** Morgan & Claypool Publishers, 180p., 2012.

SARKAR, D. **Text Analytics with Python**: **A Practical Real-World Approach to Gaining Actionable Insights from Your Data**. Library of Congress Control Number, Apress, 385p., 2016.

SOUZA, M.; VIEIRA, R.; BUSETTI, D.; CHISHMAN, R. e ALVES, I. M. Construction of a Portuguese Opinion Lexicon from multiple resources. In: **8th Brazilian Symposium in Information and Human Language Technology**, 2012

APÊNDICE A – TÍTULO

Apêndice A – Exemplo de remoção de tags HTML

```
<P>Excelente post, Mariella!<P>
<P>Acho que você e o Bruno poderão e deverão ter um papel
fundamental nesta Rede, orientando-nos em relação a como
redigir um post<P>
<P>Precisaremos, certamente, de um "manual de redação"!</P>
<P>Um abraço,</P>
<P>Ricardo<BR></P>

Excelente post, Mariella! Acho que você e o Bruno poderão e deverão
ter um papel fundamental nesta Rede, orientando-nos em relação a
como redigir um post. Precisaremos, certamente, de um "manual de redação"!
Um abraço, Ricardo
```

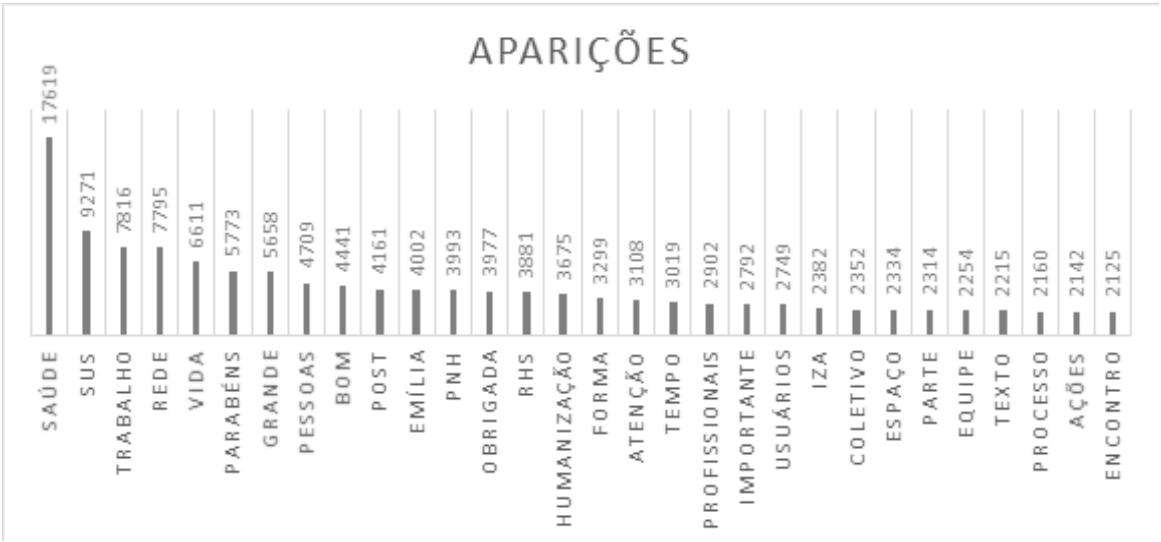
Fonte: Dados da pesquisa, 2018

Apêndice B – Exemplo de remoção de símbolos e caracteres especiais e stopwords

```
excelente post mariella acho bruno poderão deverão
papel fundamental rede orientando-nos redigir post
precisaremos certamente manual redação abraço ricardo
```

Fonte: Dados da pesquisa, 2018

Apêndice C – 30 palavras mais comuns nos comentários



Fonte: Dados da pesquisa, 2018

ARBOVIS:
Uma aplicação de visualização de dados sobre Arboviroses

ARBOVIS:
A data visualization application on Arboviruses

Talitha Alessandra Medeiros Ribeiro[1], Larissa de Sousa Moreira Gusmão[1], Damires Yluska de Souza Fernandes[1]
(1) Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), Av. Primeiro de Maio, João Pessoa - PB, medeirostalitha@gmail.com, larissamgusmao@gmail.com, damires@ifpb.edu.br.

Resumo:
O volume de informações em saúde, em especial no cenário das arboviroses, cresceu juntamente com os desafios para trabalhar com os dados de forma que análises úteis possam ser feitas e usadas para apoiar a tomada de decisões. Este artigo apresenta a Arbovis, uma aplicação para análise visual de dados sobre arboviroses no estado da Paraíba. A arbovis disponibiliza visões interativas e customizadas sobre os referidos dados, baseadas em filtros que podem ser aplicados para gerar gráficos. Os dados apresentados foram obtidos através da Secretaria de Saúde do Estado da Paraíba e do IBGE. Como resultados iniciais, foram desenvolvidos gráficos de barras, barras horizontais e de linhas. Através da visão interativa, é possível identificar quais municípios possuem maior incidência de casos das doenças, por meio do cruzamento entre quantidade de habitantes e de casos, e também em quais houve maior quantidade de casos registrados. Dessa maneira, a Arbovis apresenta impacto positivo na gestão e na prevenção de doenças e epidemias, contribuindo para estudos e análises de pesquisadores, profissionais da saúde, gestores e da população em geral.
Palavras-chave: Arboviroses. Análise de dados. Visualização de dados. Aplicação web.

Abstract:
The volume of health information, particularly in the arboviruses scenario, has grown along with the challenges of working with data in such a way that useful analyzes can be produced and used to support decision making. In this light, this paper presents Arbovis, an application for visual analysis of data on arboviruses. Arbovis provides interactive and customized views of the data, based on filters that can be applied to generate graphs. The data presented were obtained from the Department of Health of the State of Paraíba and also from the IBGE. As initial results, bar graphs, horizontal bars and lines were developed. Through an interactive view, it is possible to identify which cities have a higher incidence of arboviruses cases. This is obtained by crossing and integrating data from some data sources with registered cases. Thereby, Arbovis provides a positive impact on the management and prevention of diseases and epidemics, thus contributing to studies and analyzes of researchers, health professionals, managers and the population in general.
Keywords: Arboviruses. Data analytics. Data visualization. Web application.

1 INTRODUÇÃO

A demanda por informações em saúde cresce a cada dia juntamente com os desafios para que sua utilização possa trazer resultados úteis à população. Em especial, o cenário associado às doenças transmitidas pelo mosquito *Aedes aegypti* (arboviroses) tem causado necessidades específicas de estudos e definições de políticas públicas. Diante disso, diversos órgãos institucionais, além de pesquisadores e profissionais de saúde, trabalham gerando dados sobre arboviroses.

À medida que os dados se tornam mais volumosos, com estruturas heterogêneas e com alta velocidade de produção, como é o caso de dados em massa, ou *big data* (GANDOMI; HAIDER, 2015), a complexidade nas tarefas de análise de dados aumenta. Isso se deve a fatores como o próprio volume de dados, a quantidade de atributos a serem gerenciados e interpretados (variedade de dados) e também a aspectos da qualidade dos dados (THABET; SOOMRO, 2015). Nesse sentido, estratégias de ETL (Extração, Transformação e Carga) e de visualização da informação podem ajudar (BANSAL; KAGEMANN, 2015).

Estratégias de ETL referem-se a processos que permitem extrair dados de diferentes fontes, transformá-los para atender as necessidades de padronização e uso, e carregá-los em repositórios de destino. Com os dados extraídos, o próximo passo é visualizá-los. A visualização da informação pode ser definida como o uso de representações visuais interativas de dados abstratos para amplificar a cognição (SHNEIDERMAN; PLAISANT; HESSE, 2013). Está associada à transformação de dados abstratos em, por exemplo, gráficos e imagens, denominados visualizações (VIEIRA; CORREA, 2011).

Nesse contexto, surge o termo "análise visual de dados" que busca integrar técnicas de visualização da informação com as de análise de dados. Mais precisamente, o termo *Visual Analytics* ou Análise Visual foi introduzido por Thomas e Cook (2005) quando se percebeu que a representação visual utilizando apenas técnicas de visualização de informação não era suficiente para viabilizar a análise de dados em massa (CARMO, 2015). Técnicas de interação e manipulação direta foram associadas à atividade de análise de modo que o usuário (que tem o papel de analista de dados) pudesse ser o centro e o coordenador da atividade. Estratégias que incluem a aplicação de filtros nos dados, operações de seleção de objetos visuais, navegação, exploração por zoom, construção de representações visuais podem ampliar as possibilidades das tarefas analíticas (FEKETE, 2013). Tais estratégias podem facilitar a personalização das tarefas de análises de dados e torná-las mais intuitivas.

Nessa perspectiva, este artigo propõe a aplicação Arbovis[1], que visa apresentar dados sobre arboviroses, no estado da Paraíba, por meio de visualizações interativas. A versão apresentada neste trabalho é uma extensão da aplicação de Santiago et al. (2018), que não incluía opções de interatividade e análises visuais construídas dinamicamente. Como resultado, o ambiente de análise de dados se torna mais rico, interativo e as informações geradas são mais compreensíveis.

2 OBJETIVOS

O objetivo da Arbovis é mostrar visões significativas sobre arboviroses, distribuindo filtros de apresentação de forma que o usuário possa montar visualizações específicas dos dados de acordo

1 http://arboviz.siderg.com.br/

com suas necessidades. Desse modo, as visualizações pretendem contribuir para análise visual de dados sobre arboviroses por meio de recursos de interatividade como filtros nos dados, operações de seleção de objetos, navegação e exploração em representações gráficas diferentes.

3 TRABALHOS RELACIONADOS

Alguns exemplos de aplicações serviram como base para compor a abordagem proposta por este trabalho. A VisPublica (RIBEIRO et al., 2012) apresenta técnicas de visualização da informação (InfoVis) que são aplicadas para facilitar a transparência de dados públicos. Esta aplicação permite que os usuários utilizem as técnicas e tecnologias mostradas no catálogo para gerar as visualizações dos seus próprios dados.

A Vis-Saúde (KAIESKI, 2014) é um sistema de código aberto, no qual os registros de saúde pública são usados para fornecer análises. A aplicação propõe uma metodologia de análise e visualização da correlação entre as variáveis associadas à ocorrência de uma enfermidade em uma região.

A DataViva[2] é uma plataforma de pesquisa com dados abertos que tem como objetivo fornecer dados estatísticos sobre saúde, transporte público, economia, educação e outros domínios de dados associados a várias cidades brasileiras. Recentemente, a DataViva incluiu dados fornecidos pelo DATASUS, o órgão público da saúde do governo brasileiro.

4 PROCEDIMENTOS METODOLÓGICOS

Para viabilização do desenvolvimento da aplicação, foram realizadas reuniões com gestores da Secretaria de Saúde do Estado da Paraíba. O levantamento dos requisitos foi feito por meio de entrevistas, em que foram identificados os aspectos que os gestores consideram relevantes para a análise dos dados e que gostariam de visualizar na aplicação.

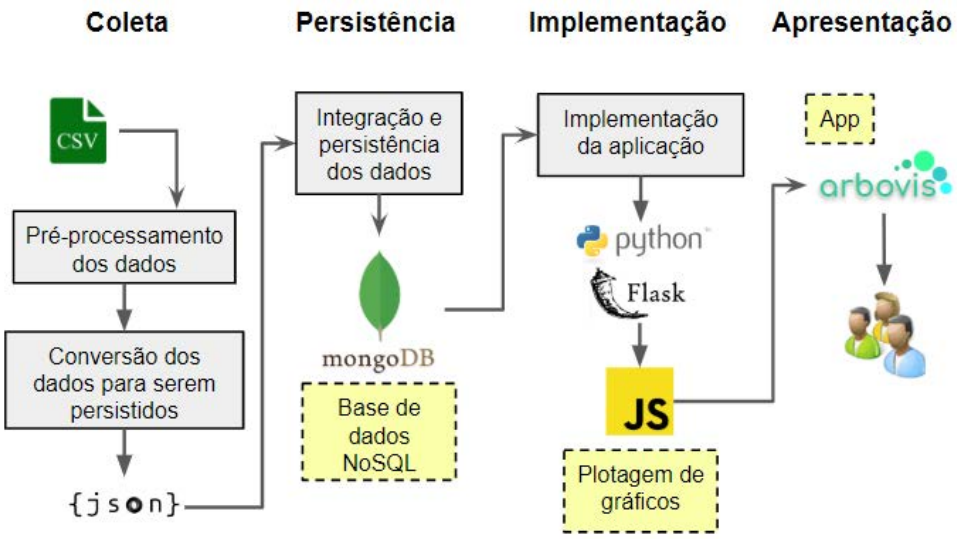
Por meio dessas reuniões, foram obtidas planilhas em formato CSV com os dados sobre as arboviroses. Os municípios e os dados sobre o quantitativo de habitantes foram obtidos através do IBGE[3].

Para embasamento técnico, foram realizados testes com tipos de gráficos diferentes utilizando as linguagens de programação Python[4] e JavaScript[5], o objetivo era identificar qual apresentação e tecnologia seriam mais eficazes para o desenvolvimento da análise interativa visual.

A Figura 1 mostra o processo empregado para o desenvolvimento da aplicação.

2 http://dataviva.info/pt/
3 https://www.ibge.gov.br/geociencias-novoportal/organizacao-do-territorio/estrutura-territorial/15761-areas-dos-municipios.html?=&t=downloads
4 https://www.python.org/
5 https://www.javascript.com/

Figura 1: Visão geral do processo para implementação.



Fonte: Os autores.

Como mostra a Figura 1, para a implementação da aplicação, primeiramente foi feito o pré-processamento dos dados obtidos. Após isso, os dados foram convertidos para o formato JSON[6], foram integrados e persistidos na base de dados MongoDB[7]. Em seguida, a aplicação foi implementada utilizando Python juntamente com o microframework Flask[8]. A integração entre o Python e o MongoDB foi feita através da distribuição PyMongo[9]. Por fim, foi iniciada a plotagem dos gráficos com a biblioteca JavaScript Chart.js[10].

5 ARBOVIS: ABORDAGEM E IMPLEMENTAÇÃO

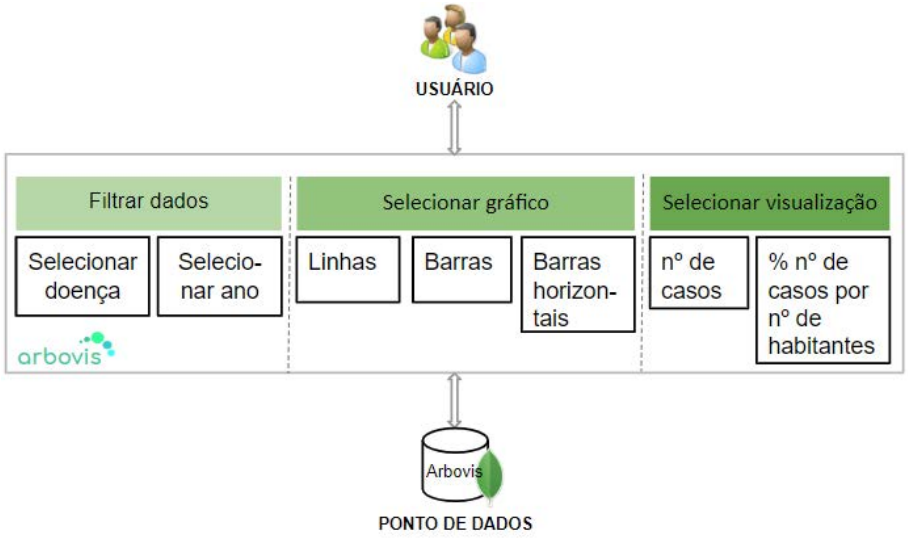
A Arbovis foi desenvolvida com foco em apresentar informações sobre as arboviroses de maneira clara e interativa para gestores, profissionais de saúde e para a população em geral.

5.1 Abordagem e funcionalidades

As principais funcionalidades da aplicação são i) filtrar dados, ii) selecionar gráfico e iii) selecionar visualização. De acordo com a pesquisa montada pelo usuário, a consulta é realizada no banco de dados e o resultado é retornado para o usuário. Uma visão geral da abordagem da Arbovis é apresentada na Figura 2.

6 http://www.json.org/
7 https://www.mongodb.com/
8 http://flask.pocoo.org/
9 https://api.mongodb.com/python/current/
10 https://www.chartjs.org/

Figura 2: Abordagem da Arbovis.



Fonte: Os autores.

5.2 Implementação e cenário de uso

Com o objetivo de facilitar o acesso as visualizações, a seção de indicadores foi evidenciada no painel da página principal, na barra de navegação e no rodapé. A Figura 3 mostra a tela principal da aplicação, com destaque para a opção de montar o gráfico de acordo com os filtros disponibilizados.

Figura 3: Página principal da Arbovis.

Fonte: Arbovis.

Além das funcionalidades principais, o usuário também é informado sobre a sintomatologia das arboviroses, há também uma breve descrição sobre as doenças (Figura 4).

Figura 4: Página principal da Arbovis.

Fonte: Arbovis.

A página de indicadores contém a visualização dos dados sobre as doenças. Esta seção é composta por filtros que o usuário pode selecionar de acordo com o objetivo da sua pesquisa, as opções selecionadas geram o gráfico dinamicamente (Apêndice A). Os filtros apresentados são i) Selecionar doença; ii) Selecionar ano; iii) Selecionar tipo de gráfico; iv) Selecionar tipo de visualização.

Considerando um cenário no qual no filtro de doença é selecionada a chikungunya, no filtro de ano de ocorrência é selecionado o ano de 2017, o gráfico apresentado é o de linhas, que por padrão já é selecionado, e número de casos absoluto (Apêndice B), é possível perceber que o município mais representativo, com 4556 casos registrados, é João Pessoa. Entretanto, se for aplicada a visualização por porcentagem do número de casos por número de habitantes, como mostrado no Apêndice C, o município de São Bentinho se destaca entre os outros, com 5.35% de casos por número de habitantes.

Outro cenário pode ser visto no Apêndice D, em que são selecionados a doença dengue e o ano de 2016 com gráfico de barras e número de casos absolutos registrados por município. Neste contexto, o município com mais registros é João Pessoa, com 7701 casos. Já no cenário de porcentagens do número de casos pelo número de habitantes, como mostrado no Apêndice E, o município mais significativo é Juripiranga, com 27.58% de casos por número de habitantes.

Através das visões apresentadas, é possível identificar quais são os municípios que necessitam de maior atenção do governo, por possuírem maior incidência de casos das doenças. Dessa forma, as visualizações disponibilizadas pela Arbovis possibilitam ações mais assertivas por parte dos gestores.

5.3 Avaliação da Arbovis

Com o objetivo de analisar aspectos da Arbovis como a usabilidade na utilização e apresentação dos indicadores, foi realizada uma avaliação inicial com um gestor da Secretaria de Saúde da Paraíba.

No aspecto da utilização da Arbovis, o gestor considerou a aplicação fácil de usar e também indicou que ela o ajudou a encontrar informações importantes sobre as arboviroses.

Outro aspecto apontado é a apresentação dos indicadores, em que o usuário considerou os filtros adequados para a realização das consultas e que tornaram a aplicação interativa. Além disso, informou que a aplicação apresenta praticidade nos resultados apresentados. Como sugestão de melhoria, foi sugerido acrescentar índices de infestação predial por município junto às taxas de incidência dos agravos, estes são dados que serão futuramente integrados à base existente.

6 CONSIDERAÇÕES FINAIS

A Arbovis é pautada em um problema social e possibilita o uso de dados reais em um cenário de demandas concretas a esses dados. Os resultados obtidos com esta aplicação possuem impacto especial na gestão em relação a prevenção de doenças e epidemias causadas pelo *Aedes aegypti*.

Pesquisadores, professores, gestores ou quaisquer interessados poderão consultar os dados de forma integrada e analisá-los de acordo com suas necessidades. Para os gestores, a aplicação é significativa pois os resultados obtidos através das visões customizadas podem contribuir para a tomada de decisão. Entretanto, para expandir as possibilidades de análise dos gestores esperamos incluir em trabalhos futuros o uso de mapas e averiguar a eficácia dessa visualização na gestão.

Também podem ser identificadas necessidades específicas de diferentes perfis, para contribuir em maior profundidade em seus estudos e análises. Um perfil que se destaca para ser explorado é o profissional de saúde.

REFERÊNCIAS

BANSAL, S.; KAGEMANN, S. Integrating Big Data: A Semantic Extract-Transform-Load Framework. **Computer**, vol.48, n.3, p. 42-50, 2015.

CARMO, A.. **Exploração e análise de dados coletados pelo Sistema Integrado de Monitoramento Ambiental utilizando técnicas de Visual Analytics**. 2015. Dissertação de Mestrado. Faculdade de Ciências e Tecnologia (FCT), Presidente Prudente, 2015. Disponível em: <https://repositorio.unesp.br/handle/11449/138391>. Acesso em: 04 ago. 2018.

FEKETE, J. Visual Analytics Infrastructures: From Data Management to Exploration. **Computer**, Institute of Electrical and Electronics Engineers, Visual Analytics: Seeking the Unknown, vol.46, n.7, p.22-29, 2013.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, Ontario, v.35, p.137-144, 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0268401214001066>. Acesso em: 05 out. 2018.

KAIESKI, N. **Vis-Saúde - Uma metodologia para visualização e análise de dados de saúde pública**. 2014. Dissertação de Mestrado (UNISINOS), São Leopoldo, 2014. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/3619>. Acesso em: 04 ago. 2018.

RIBEIRO, F. et al. VisPublica: uma proposta para aprimorar a transparência de dados públicos. In: **VIII Simpósio Brasileiro de Sistemas de Informação**, 2012, São Paulo. p. 480-491.

SANTIAGO, M., ALENCAR A., FERNANDES A., SOUZA, D. (2018). A Semantic-based Approach for Facilitating Arbovirus Data Usage. 20º International Conference on Enterprise Information Systems (ICEIS), Funchal, 2018.

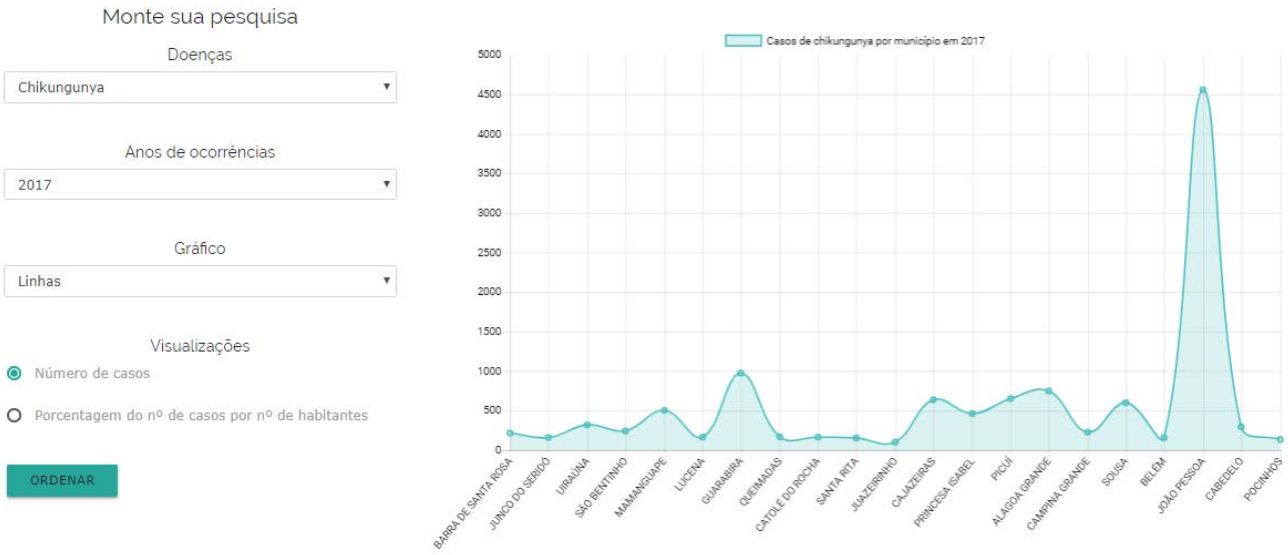
SHNEIDERMAN, B., PLAISANT, C., HESSE, B. Improving health and healthcare with interactive visualization methods. **Computer**, vol. 46, n. 5, p. 58-66, 2013.

THABET, N.; SOOMRO, T. Big Data Challenges. **Journal of Computer Engineering & Information Technology**. Dubai, 2015. Disponível em: <https://www.researchgate.net/publication/282281171_Big_Data_Challenges>. Acesso em: 04 out. 2018.

THOMAS, J. J.; COOK, K. A. (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics. [S.l.]: National Visualization and Analytics Ctr. Paperback. ISBN 0769523234.

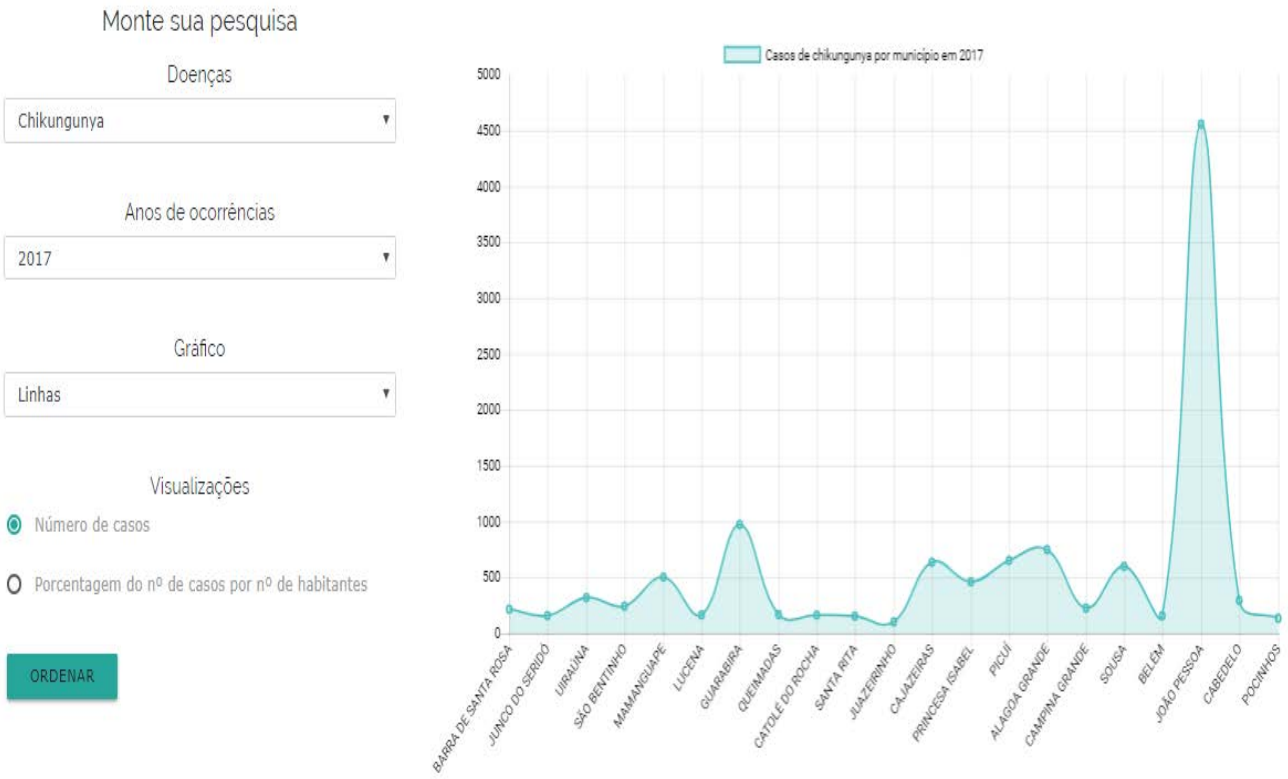
VIEIRA, J., CORREA, R. Visualização da Informação na construção de interfaces amigáveis para Sistemas de Recuperação de Informação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 16, n. 32, p. 73-93, out. 2011. ISSN 1518-2924. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2011v16n32p73>. Acesso em: 04 ago. 2018.

Apêndice A - Filtros selecionados para visualização customizada.



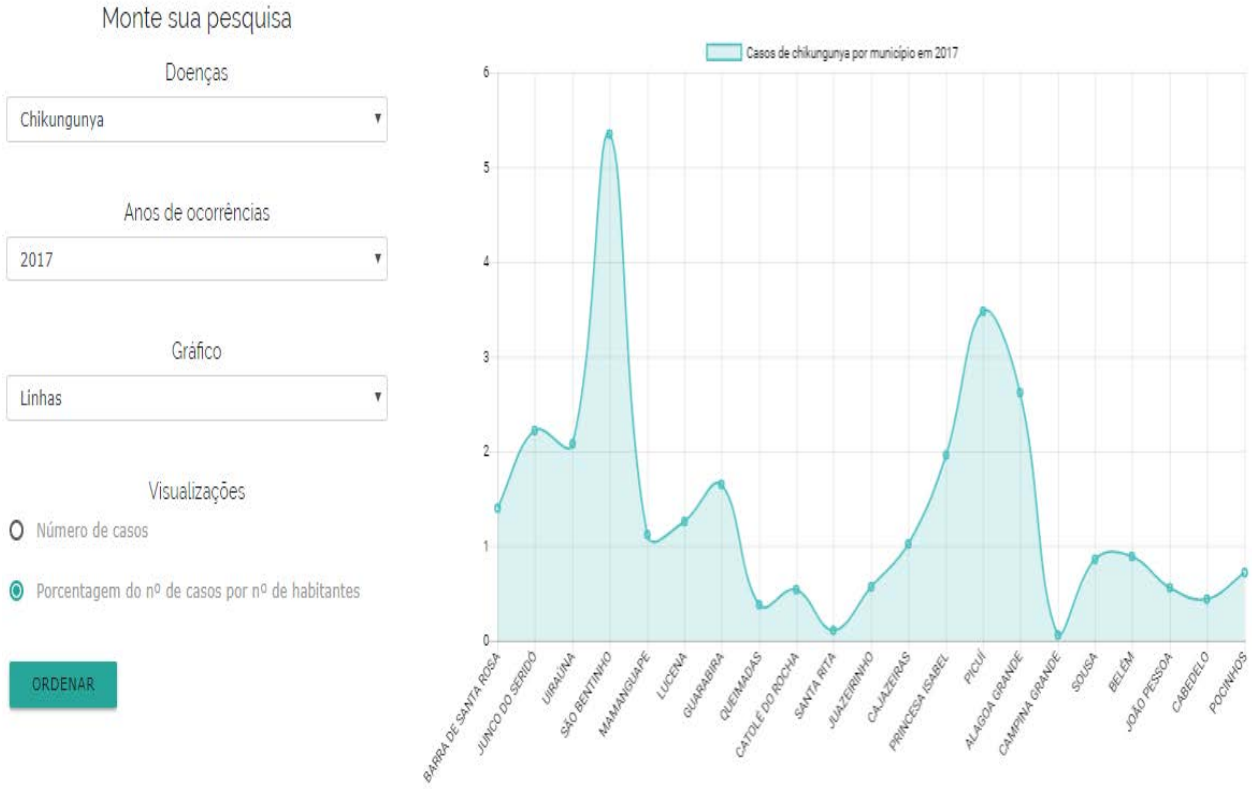
Fonte: Arbovis.

Apêndice B - Gráfico de linhas com casos de chikungunya por município em 2017 com número de casos total.



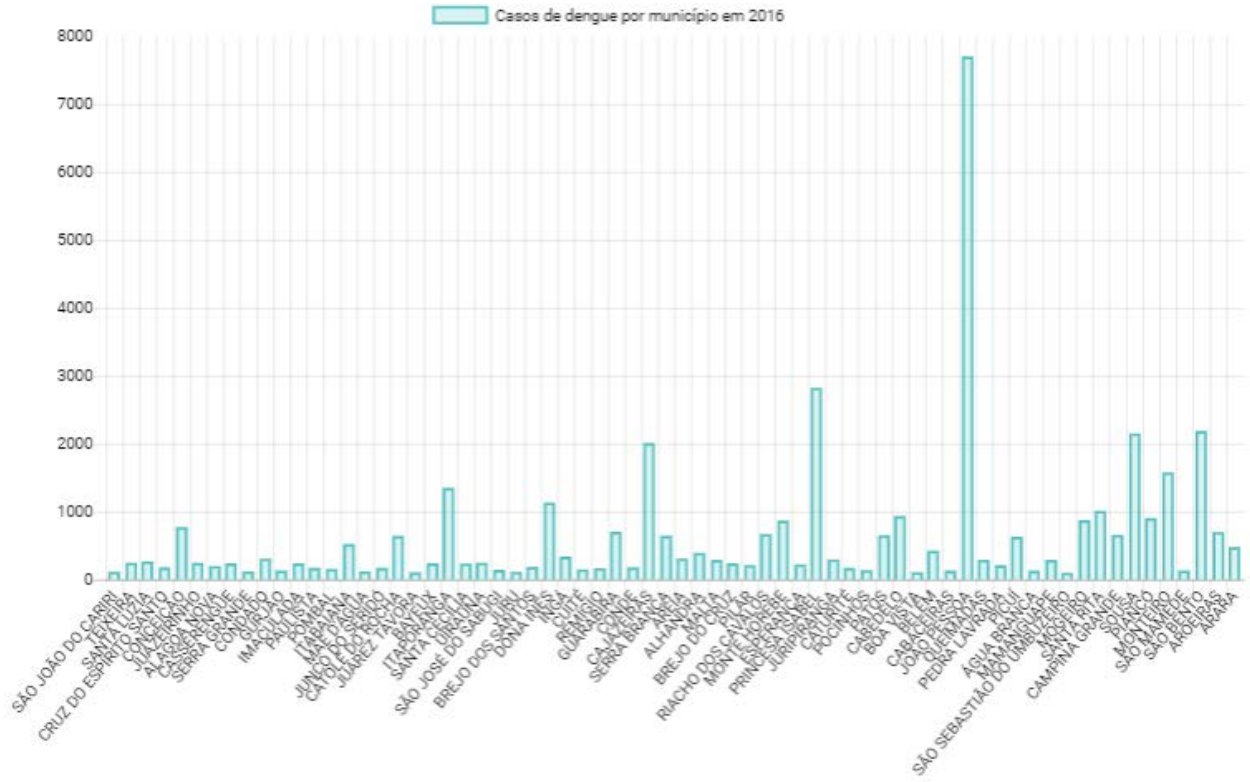
Fonte: Arbovis.

Apêndice C - Gráfico de linhas com casos de chikungunya por município em 2017 com porcentagem de casos por número de habitantes.



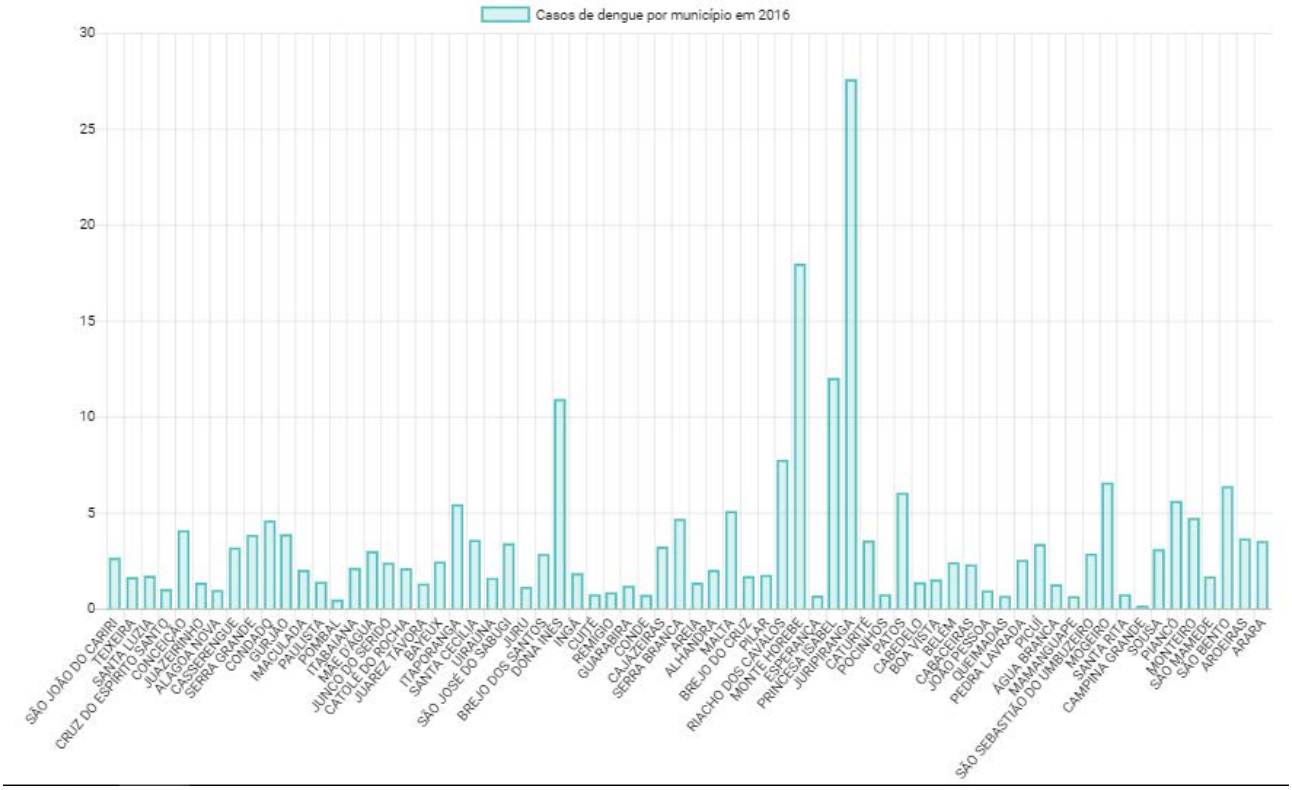
Fonte: Arbovis.

Apêndice D - Gráfico de barras com casos de dengue por município em 2016 com número de casos total.



Fonte: Arbovis.

Apêndice E - Gráfico de barras com casos de dengue por município em 2016 com porcentagem de casos por número de habitantes.



Fonte: Arbovis.

AS LICENÇAS CREATIVE COMMONS E SOFTWARE LIVRE COMO FORMAS DE PROTEÇÃO NA SOCIEDADE DA INFORMAÇÃO COMPARTILHADA

THE CREATIVE COMMONS AND FREE SOFTWARE LICENSES AS FORMS OF PROTECTION IN THE SHARED INFORMATION SOCIETY.

Maria Carolina Zanini Ferreira[1], Clarissa Stefani Teixeira[2], Enrique Muriel-Torrado[3]
(1) Programa de Pós-Graduação em Propriedade Intelectual e Transferência de Tecnologia para a Inovação (PROFNIT),Universidade Federal de Santa Catarina (UFSC), Campus Universitario Reitor João David Ferreira Lima, Florianópolis – SC, CEP: 88040-900, mariacarolina.zanini@gmail.com.
(2) Programa de Pós- Graduação em Engenharia e Gestão do Conhecimento (EGC), Universidade Federal de Santa Catarina (UFSC), Campus Universitario Reitor João David Ferreira Lima, Florianópolis – SC, CEP: 88040-900, clastefani@gmail.com.
(3)Programa de Pós-graduação em Ciência da Informação (PGCIN), Universidade Federal de Santa Catarina (UFSC), Campus Universitario Reitor João David Ferreira Lima, Florianópolis – SC, CEP: 88040-900, enriquemuriel@gmail.com.

Resumo:
O presente estudo possui a finalidade de entender de que formas as licenças creative commons e software livre são uma alternativa de proteção ao capital intelectual em uma sociedade da informação compartilhada. Por meio de um estudo bibliográfico de caráter descritivo exploratório, buscou contextualizar a sociedade da informação que trouxe como consequência um mundo mais conectado e principalmente colaborativo. Logo, discussões começaram a ser realizadas acerca da proteção da propriedade intelectual, nas quais obras intelectuais, novas tecnologias, estão sendo compartilhadas, desenvolvidas de forma coletiva e ainda distribuídas em um mundo digital, que a princípio poderia ferir os direitos que são assegurados ao autor desses ativos. Como resultado, mecanismos de licenciamento começaram a ser encarados como uma maneira que assegura os direitos do autor mas ao mesmo tempo permite o compartilhamento das obras autorais desde que sejam respeitadas algumas condições. Com isso, mostra que os direitos autorais em um mundo conectado e colaborativo pode ser revisto em padrões que não infrinjam suas próprias regras e ainda por cima contribuir para que uma forma mais segura e colaborativa possa se dar em um mundo digital com uma vasta rede de dados sendo usadas, reusadas e distribuídas.
Palavras-chave: Licenças. Creative Commons. Software Livre. Sociedade da Informação.

Abstract:
The purpose of this study is to understand the ways in which creative commons and free software licenses are an alternative to protect intellectual capital in a shared information society. Through a bibliographic study characterized as exploratory descriptive, it sought to contextualize the information society that brought as a consequence, a more connected and mainly collaborative world. Therefore, discussions began to take place about protection of intellectual property, in which intellectual works, new technologies, are being shared, developed collectively and even distributed in a digital world, which in principle could violate the copyright. As a result, licensing mechanisms have begun to be seen as a security option to copyright, but at the same time allowing the sharing of copyright works, provided that certain conditions are respected. In this way, it shows that copyright in a connected and collaborative world can be reviewed in patterns that do not violate its own rules and, moreover, contribute to a more secure and collaborative way in a digital world with a vast network of data being used, reused and distributed.
Keywords: Licenses. Creative Commons. Free Software. Information Society.

1 INTRODUÇÃO

A globalização trouxe a informação e o conhecimento como os principais ativos do século, onde as práticas sociais estão sendo examinadas e (re)formadas constantemente. A entrada de informações que alteram não somente o indivíduo ou um grupo específico, mas toda a comunidade global é impulsionada cada vez mais pela evolução crescente das tecnologias de informação e comunicação (TICs), onde tudo está conectado e interligado (GIDDENS, 1990; KAWAMURA, 2007).

Estudos recentes sobre esse ativoque é a informação e a sua transformação em conhecimento de alto valor agregado como vetor para a inovação, são ressaltados por Carmeli, Gelbard e Reiter-Palmon (2013) como um processo essencialmente colaborativo e comunicativo, fruto do fluxo contínuo de informação com o entorno produtivo e a sociedade, criando ambientes que potencializam a criatividade e novos mecanismos de compartilhamento.

Isso naturalmente leva a uma série de perguntas sobre as questões e riscos do uso sobre o compartilhamento de obras intelectuais, novas tecnologias como software que atualmente trazem consequências na esfera dos direitos autorais, arcabouço legal que assegura ao autor a proteção de direitos sobre a sua obra automaticamente obtida com a sua criação (BÉRANGER; RAVIX, 2013, EVENS, 2015).

O resultado dessas discussões levou ao surgimento de novos mecanismos legais que buscam assegurar os direitos intelectuais ao mesmo tempo em que permitem acesso público e o uso da informação contidas nas obras, sendo conhecidos hoje como licenças de software livre e *creative commons* (YANG; MASKUS, 1999; XUE, 2014).

Logo, por se tratar de um tema atual e que ainda necessita de maior compreensão, o presente trabalho busca contribuir para a discussão acerca da proteção intelectual frente a uma sociedade da informação de caráter colaborativo.

Pretende-se, portanto, não exaurir o tema, mas apresentar uma alternativa para agregar a discussão existente sobre a propriedade intelectual e o compartilhamento da informação, temas que regularmente são tratados como antagônicos (YANG; MASKUS, 1999; XUE, 2014).

2 OBJETIVOS

A proposta do presente estudo é compreender como as licenças creative commons e software livre podem ser uma opão de proteção em uma sociedade da informação de caráter colaborativo.

3 PROCEDIMENTOS METODOLÓGICOS

O procedimento metodológico da pesquisa se caracteriza como descritiva exploratória, a partir de uma contextualização geral que buscou evidenciar a sociedade da informação em um mundo globalizado onde novos paradigmas emergem frente a uma dinâmica colaborativa derivada do

desenvolvimento das TICs. Dentre as diversas discussões oriundas desse cenário, serão ressaltadas as questões relacionadas a propriedade intelectual, mais precisamente no campo dos direitos autorais, onde novos mecanismos legais de proteção e licenciamento denominadas como licenças creative commons e software livre são apresentadas (MARCONI; LAKATOS, 2003).

Logo, foram utilizadas técnicas padronizadas de coleta de dados e observação sistemática, utilizando fontes e dados secundários obtidos por meio de pesquisa documental provindos da legislação brasileira relacionado a direitos autorais, bancos de dados tais como scopus, web of science, filtrando estudos que possuíssem informações referentes aos temas estudados (sociedade da informação, direitos autorais, creative commons e software livre) assim como fontes provenientes de organizações e instituições de ensino e pesquisa que pudessem auxiliar a proposta do referente tema (PRODANOV; FREITAS, 2013).

De modo a atender o objetivo, o estudo foi compreendido em três fases: (i) contextualizar a sociedade da informação e suas as implicações contidas no âmbito dos direitos autorais e; (ii) apresentar as licenças de software livre e creative commons como forma de proteção da informação em mundo colaborativo.

4 RESULTADOS

O mundo contemporâneo temmudado rapidamente a realidade, orientado tecnologicamente para um ambiente totalmente virtual de armazenamento de dados e informações. Tal percepção do "ciberespaço" tende a desregulamentar barreiras de acesso a informação e conhecimento por meio de um ambiente altamente colaborativo e compartilhado (LEVIN, 2014).

Apesar do tema não ser novo, as formas de compartilhamento de conhecimento e informação mudaram drasticamente, e com isso trouxeram diversos questionamentos para a sociedade global, fruto da quantidade crescente de informações disponíveis, da sua transformação em um ativo especial, a efetiva combinação e a utilização das demais informações e da tamanha complexidade que essas relações acabaram trazendo na conjuntura prática da sociedade da informação em um mundo globalizado (SOUSA; AMARAL, 2012).

Com a expansão da internet em meados dos anos 90, iniciou-se uma dinâmica de compartilhamento de conhecimento e sobretudo acesso universal à recursos informacionais, potencializando um grande canal de comunicação no que concerne a produção científica por meio revistas eletrônicas, repositórios acadêmicos, bibliotecas digitais, entre outros. Esse cenário permitiu que novos recursos de conhecimento fossem disponibilizados para acesso público ou em consumo público em larga escala, geralmente sem qualquer impedimento de taxa de assinatura ou taxas de acesso, originando assim o movimento de acesso aberto (UNESCO, 2015).

Essa reprodução e a distribuição de obras digitais possuem um capital intelectual que é protegido pela lei dos direitos do autor, acarretando questionamentos e mudanças sobre a efetivação da tutela jurídica que acaba sendo afetada por essas novas formas de interação e tecnologias (ARAYA; VIDOTTI, 2011).

O direito autoral faz parte do que se conhece hoje como propriedade intelectual, isto é, uma proteção dada a todas as criações decorrentes do espírito humano de caráter científico, literário, artístico ou industrial (DUARTE; PEREIRA, 2009).

Pode-se dizer que o direito autoral tem como objetivo conceder proteção legal a uma vasta variedade de obras literárias, como é o caso de livros, revistas, jornais, música, pinturas, fotografias, esculturas, entre outros. Essa proteção tem uma duração limitada o qual permite que o autor explore patrimonialmente a sua obra, por qualquer forma "direta ou indireta" (INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL, 2013).

No Brasil a legislação que regula os direitos do autor é a Lei 9610/98. A proteção da obra intelectual permanece durante toda a vida do autor sendo transmissível por mais de 70 anos e independe de registo. A sua redistribuição depende inteiramente de autorização do autor da obra (BRASIL, 1998). A legislação brasileira poderia ser considerada, pela sua data de promulgação e pelo seu conteúdo, como uma legislação pré-internet que não contempla muitas das interações e transformações digitais ocorridas ao longo dos anos.

Com isso, discussões acerca da disponibilização e compartilhamento de obras de caráter intelectual para terceiros no meio virtual, principalmente no que se concerne a "novas" obras intelectuais, como é o caso do software, começam a entrar na pauta da proteção intelectual.

De acordo com Muriel-Torrado e Fernández-Molina (2014), o desajuste da legislação vigente e as mudanças proporcionadas pelo ambiente digital, ainda é grande e impedem que as atividades possam decorrer com normalidade e segurança devida. Muitas pessoas ainda não sabem que ao compartilhar uma obra, automaticamente estará refringindo os direitos autorais daquele que o possui, salvo em caso de autorização prévia ou licenciamento.

Em decorrência de tais discussões, iniciativas como as licenças creative commons e software livre (caracterizadas como copyleft) começaram a estender a liberdade de compartilhar ou reutilizar ativos como obras intelectuais e softwares por meio de proteção de direitos autorais (UNESCO, 2015).

O movimento do software livre foi iniciado com Richard Stallman que trazia em seu discurso o livre acesso ao software em uma época na qual algumas empresas, como a Microsoft, começaram a pressionar as autoridades governamentais, para torná-lo passível de patente (STALLMAN, 2015).

O pontapé inicial foi o desenvolvimento de um sistema operacional completamente livre em 1983, chamado GNU que fora compartilhado por uma licença chamada general public license (GPL), que usava o próprio regime dos direitos do autor para restringir a capacidade de outras pessoas de usá-lo, distribuí-lo, copiá-lo ou desenvolver outro sistema a partir deste,) a qual posteriormente, tornou-se a principal referência para a criação de diversas licenças de código aberto e a formação da Free Software Foundation, organização que prega o livre acesso ao software com novas licenças de distribuição e compartilhamento desde que sejam respeitadas as quatro liberdades principais; a de execução; modificação; redistribuição e distribuição de versões modificadas (STALLMAN, 2015).

Porém, a preocupação sobre questões autorais não estava somente relacionada ao software. As obras intelectuais maciçamente compartilhadas, como já mencionadas, em bibliotecas digitais, canais de comunicação e produção online começaram a levantar questionamentos sobre a distribuição de obras autorais. Por conta disso em 2002, inspirado no movimento do software livre, Lawrence Lessig elaborou licenças públicas padronizadas, chamadas creative commons, estabelecendo direitos concedidos a obras intelectuais entre elas livros, artigos, filmes, músicas, entre outros. Com isso, pretendia-se tornar mais fácil a organização, o acesso, o compartilhamento e modificação dessas obras em rede (BRANCO, BRITTO, 2013; LIU, 2014).

Ao todo, são seis licenças creative commons padronizadas, com variação da mais permissiva a mais restritiva, onde a primeira possibilita a distribuição, remixagem, adaptação ou criação de obras derivadas, mesmo que seja para fins comerciais, contando que seja dado crédito pela criação original. Já a segunda, permite que terceiros possam remixar, adaptar e criar obras derivadas ainda que para fins comerciais, contanto que o crédito seja atribuído ao autor e que essas obras sejam licenciadas sob os mesmos termos. Indo para a terceira, é permitido a redistribuição e o uso para fins comerciais desde que a obra seja redistribuída sem modificações e completa, creditando o autor. A quarta, autoriza a criação de obras derivadas sobre a obra licenciada assim como remixagem e adaptação, porém proíbe o uso tanto da obra como novas obras derivadas fins comerciais. A próxima licença possibilita que outros também remixem, adaptem e criem obras derivadas sobre a obra original sob a condição que não seja comercializado, atribuam crédito ao autor e que licenciem as novas criações da mesma maneira. Por último, a mais restritiva, permitindo que outros apenas façam download e compartilhem as obras licenciadas, entretanto, não poderá modificar a obra de nenhuma forma, muito menos utilizá-la para fins comerciais (BRANCO; BRITTO, 2013).

Esses mecanismos trouxeram importantes reflexões sobre a necessidade de oferecer um arcabouço legal por meio de um licenciamento, que pudessem expressar o compartilhamento sem negligenciar os direitos do autor. Essas licenças o ferecem uma convergência entre a proteção e a rede colaborativa incluindo ferramentas necessárias para desenvolver e distribuir programas e obras sem fronteiras ao mesmo tempo que obriga que todos os licenciados façam referência ao autor da obra e alguns casos até utilizem o mesmo modelo de licenciamento nas suas redistribuições inclusive em suas versões derivadas (MONIZ; CERDEIRA, 2004; LEMOS; BRANCO JÚNIOR, 2006).

Ao invés de permitir que um trabalho seja totalmente disponibilizado em domínio público, onde não existe nenhuma restrição de direitos imposta, podendo correr o risco inclusive de ter sua obra ou derivação desta protegida posteriormente por questões autorais, essas licenças permitem ao autor, impor algumas restrições aqueles que queiram utilizar seu trabalho mas que seja também permitido o compartilhamento e distribuição destes.

Apesar de já haver estudos e pesquisas apontando para modificações na própria lei dos direitos do autor, tal pauta ainda encontra-se em discussão, sendo que enquanto isso, a utilização das licenças vem crescendo ao longo dos últimos anos (ROSA, 2014). De acordo com os dados da Federação das Indústrias do Estado de Santa Catarina - FIESC ([201-]) uma das grandes macrotendências no setor da tecnologia da informação e comunicação atual são as plataformas abertas de software, onde, apenas no Brasil, 73% das empresas brasileiras utilizam softwares livres. Ainda, segundo dados da própria comunidade Creative Commons, ([201-]), mais de 1,6 bilhões de obras intelectuais vem sendo compartilhadas por meio dessas licenças. Perceber a ascensão do tema é essencial para que possa não somente abrir maiores estudos e compreensão, mas também como utilizá-las de forma adequada e responsável (CHAVES, PEREZ, CHAVES, 2013). Espera-se, portanto, que ao expor a discussão, estudos mais aprofundados sejam realizados futuramente.

5 CONCLUSÃO OU CONSIDERAÇÕES FINAIS

Em uma sociedade da informação altamente colaborativa, percebe-se que as pessoas não trabalham mais exclusivamente sozinhas. Há autores que escrevem livros contando com a contribuição

dos leitores, usuários da internet que criam diariamente obras derivadas e programas de computador de forma coletiva e que ao mesmo tempo comercializam seus produtos oriundos desses ativos.

Nesse sentido, o estabelecimento de movimentos como software livre e creative commons permite que por meio de licenças permissivas seja criado um mecanismo flexível que ao mesmo tempo que facilita o compartilhamento, a reutilização e a derivação de um conteúdo intelectual em domínios profissionais e a comunidade como um todo, traz consigo o direito de autor que defende algumas regras que devem ser respeitadas, como a referência de seu trabalho a obras de terceiros, assim como em alguns casos, a obrigação de compartilhar a obra nos mesmos formatos licenciados.

Enquanto ainda não são definidas políticas mais efetivas em relação a proteção de direitos autorais no meio cibernético, as licenças vêm como uma alternativa de reflexão de acesso,modificação e distribuição de ativos de altíssimo valor agregado protegidos por um mecanismo legal já estabelecido.

REFERÊNCIAS

ARAYA, Elizabeth Roxana Mass; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Web colaborativa: inovação na criação, proteção e uso legal de informação. **DataGramaZero - Revista de Ciência da Informação**, Rio de Janeiro, v.12 n.4, ago. 2011.

BÉRANGER, Jérôme; RAVIX, Valéry. Une société consumériste d'information orientée vers une infosphere éthique. **La Revue canadienne des sciences de l'information et de bibliothéconomie**, Toronto, v. 37. n. 4, p. 253-266, 2013. Disponível em: <http://www.adel-label.com/wp-content/uploads/2016/05/24eme-article-beranger.pdf>

BRANCO, Sérgio; BRITTO, Walter. **O que é Creative Commons? Novos Modelos de direito autoral em um mundo maiscriativo**. Rio de Janeiro: Editora FGVm 2013. 176p. Disponível em: <http://bibliotecadigital.fgv.br/dspace/bitstrea m/handle/10438/11461/O%20que%20%C3%A9%20Creative%20Commons.pdf>. Acesso em: 29 mar. 2018.

BRASIL. Lei n. 9.609, de 19 de fevereiro de 1998b. Dispõe sobre a proteção da propriedade intelectual de programa de computador, sua comercialização no País, e dá outras providências. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 19 fev. 19 fev. 1998. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L96 09.htm>. Acesso em: 17 fev. 2018.

CARMELI, Abraham; GELBARD, Roy; REITER-PALMON, Roni. Leadership Creative Problem – Solving, Capacity and Creative Performance: The importance of knowledge sharing. **Human Resource Managment,** v. 52, n.1, p. 95-122, jan/fev. 2013. Disponível em: <https://en-coller.tau.ac.il/sites/nihul_en.tau.ac.il/files/RP_237_Carmeli.pdf>. Acesso em: 17 nov. 2018.

CHAVES, Flávia Gonçalvez; PEREZ, Nuria Rojo; CHAVES, Paula Gonçalves**. A importância do uso de softwares livres na formação acadêmica**. [S.l], 2013. Disponível em:<ueadsl.textolivre.pro.br/2013.1/papers/uploa d/34.pdf>. Acesso em: 03 jul. 2018.

CREATIVE COMMONS**. Our Movement around the World**. [S.l.], [201-]. Disponível em: <https://stateof.creativecommons.org/>. Acesso em: 10 jun. 2018.

DUARTE; Eliane Cordeiro de Vasconellos Garcia; PEREIRA, Edmeire Cristina. **Direito autoral: perguntas e repostas**. Curitiba: UFPR, 2009. 164 p. Disponível em: <http://www.cipead.ufpr.br/wpcontent/uploads/2015/03/LivroDireitoAutoral.pdf>. Acessoem: 20 mar. 2018.

EVENS, Tom. Creative Commons licenses in cultural heritage institutions in Flanders. **Journal of Librarianship and Information Science**, v. 48, n. 2, p. 209-217, 2015.

FIESC. Federação das Indústrias do Estado de Santa Catarina.Tendências. [S.l.]; [201-]. Disponível em: <http://portalsetorialfiesc.com.br/grafico-tendencias>. Acesso em: 15 nov. 2018.

GIDDENS, Anthony. ***The Consequences of Modenity***. Califórnia: Standford University Press, 1990. 188 p.

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL. INPI. **Expressão criativa: uma introdução ao direito de autor e aos direitos conexos para pequenas e médicas empresas**. Rio de Janeiro: INPI, 2013. 88 p.

KAWAMURA, Takaya. Managing networks of communities of practice for organizational knowledge creation: a knowledge management imperative in the era of globalization.**Annales Des Télécommunications**, v. 62, n. 7-8, p. 734-752, 2007.

LEMOS, Ronaldo; BRANCO JÚNIOR, Sérgio Vieira. Copyleft, Software Livre e Creative Commons: A Nova Feição dos Direitos Autorais e as Obras Colaborativas. **Revista de Direito Administrativo**, Rio de Janeiro, v. 243, p. 148-167, jan. 2006. Disponível em:<http://bibliotecadigital.fgv.br/ojs/index.php/rda/article/view/42557/41318>. Acesso em: 17 Nov. 2018.

LEVIN, Ilya. Cultural Trend in a Digital Society.IN: TMCE, 10, 2014, Budapest, Hungria. **Anais...** Budapeste, 2014. 21p.

LIU, Chen-Chung et al. Online knowledge sharing experience with Creative Commons. **Online Information Review**, v. 38, n. 5, p. 680-696, 2014.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. São Paulo: Atlas, 2003. 312 p. Disponível em: <https://docente.ifrn.edu.br/olivianeta/disciplin as/copy_of_historia-i/historia-ii/china-e- india>. Acesso em: 3 abr. 2018.

MONIZ, Pedro de Paranaguá; CERDEIRA, Pablo de Camargo. Copyleft e software livre: uma opção pela razão – eficiências tecnológica, econômica e social. **Revista da ABPI**, n. 70, p. 65-70, 2004.

MURIEL-TORRADO, Enrique; FÉRNADEZ-MOLINA; J. Carlos. Enseñanza digital versus derechos de autor: el papel de la biblioteca universitária em apoyo de professores y alumnos. **Revista eletrônica de Biblioteconomia e Ciência da Informação**, v. 19, n. 39, p. 205-226, jan. /abr., 2014.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de Freitas. **Metodologia do trabalho Científico: métodos e ténicas da pesquisa e do trabalho acadêmico.** Novo Hamburgo: Feedvale, 2012, 276 p. Disponível em: <http://www.feevale.br/Comum/midias/8807f0 5a-14d0-4d5b-

-b1ad-1538f3aef538/Ebook%20Metodologia%20do%20Trabalho%2 0Cientifico.pdf>. Acesso em: 25 mar. 2018.

ROSA, Flávia. O direito autoral e o acesso aberto. In: SILVA, Rubens Ribeiro Gonçalves da (Org.). **Direito autoral, propriedade intelectual e plágio**.Salvador: EDUFBA, 2014, 214p. Disponível em: <https://repositorio.ufba.br/ri/bitstream/ri/15656/3/direito_autoral_propriedade_intelectual_plagio_RI.pdf>. Acesso em: 17 de nov. 2018.

SOUSA; Antônio José Figueiredo Peva de; AMARAL, Sueli Angélica do. Impacto do Compartilhamento da Informação e do Conhecimento para o Desenvolvimento de Inovações em Grandes Organizações. **Perspectivas em Gestão & Conhecimento**, João Pessoa, v. 2, Número Especial, p. 12-26, out. 2012. Disponível em: <http://periodicos.ufpb.br/ojs2/index.php/pgc>. Acesso em: 20 março. 2018.

STALLMAN, Richard. **Free Software, Free Society: Selected Essays of Richard M. Stallman**. Free Software Foundation: 3 ed Boston, 2015. 305 p. Disponível em: <https://www.gnu.org/doc/fsfs3-hardcover.pdf>. Acesso em 12 nov. 2018.

UNESCO. **Introduction to Open Acess**. Paris: UNESCO, 2015. 82p. Disponível em: <http://unesdoc.unesco.org/images/0023/002319/231920E.pdf>. Acesso em: 15 nov. 2018.

XUE, Nie. Research on Intellectual Property Protection under network information sharing environment**. Applied Mechanics and Materials**, v. 687, p. 191-194, 2014.

YANG, Guifang; MASKUS, Keith E. Intellectual Property Rights, licensing and innovation in an endogenous product –cycle model. **Journal of International Economics**, v. 53, p. 169-187, 1999

ASPECTOS DO CICLO DE VIDA DE DADOS EM PROCESSOS DE CONSTRUÇÃO DE ONTOLOGIAS BIOMÉDICAS

ASPECTS OF THE DATA LIFECYCLE IN PROCESSES OF CONSTRUCTION OF BIOMEDICAL ONTOLOGIES

Jeanne Louize Emygdio[1], Eduardo Ribeiro Felipe[2]
(1) Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, jeanne.emygdio@gmail.com
(2) Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, erfelipe@hotmail.com

Resumo:
Tendo por foco o estudo das coisas existentes no mundo e das relações que estas estabelecem entre si, as Ontologias vêm ganhando campo especial de investigação científica por serem consideradas alternativas viáveis à organização das informações massivas e ininterruptamente alimentadas pelos seres humanos nos espaços digitais, por meio das tecnologias da informação e comunicação (TICs). As perspectivas centrais da utilização das ontologias repousam sobre as possibilidades de prover interoperabilidade semântica entre os sistemas de informação, pouco alcançada por meio dos modelos, técnicas e ferramentais usualmente aplicados. O objetivo deste artigo é descrever de que forma a gestão do ciclo de vida dos dados pode contribuir para prover maior qualidade à construção de ontologias biomédicas. A metodologia adotada foi a pesquisa qualitativa, de caráter descritivo sobre os processos de construção da Ontologia realista Obstétrica e Neonatal - OntoNEO. Os resultados encontrados comprovaram a eficiência do plano de gestão de dados como requisito de atribuição de qualidade à referida ontologia, o que a habilitou a integrar o projeto Basic Formal Ontology - BFO, padrão de referência na construção de ontologias realistas de alto nível.
Palavras-chave: Ontologias biomédicas. Ciclo de vida de dados. Plano de gestão de dados. Representação do conhecimento.

Abstract:
With the focus of the study of the existent things in the world and of the relations that these establish with each other, the Ontologies have been gaining special field of scientific investigation for being considered viable alternatives to the organization of the massive information and uninterruptedly fed by the human beings in the digital spaces, through of information and communication technologies (ICTs). The central perspectives of the use of ontologies rest on the possibilities of providing semantic interoperability among information systems, little achieved through the models, techniques and tools usually applied. The purpose of this article is to describe how the data life cycle management can contribute to provide higher quality to the construction of biomedical ontologies. The methodology adopted was the qualitative research, of a descriptive character on the construction processes of the realistic Ontology Obstetric and Neonatal - OntoNEO. The results obtained proved the efficiency of the data management plan as a requirement of quality attribution to this ontology, which enabled it to integrate the Basic Formal Ontology - BFO project, a reference standard in the construction of high level realist ontologies.
Keywords: Biomedical ontologies. Data life cycle. Data management plan. Knowledge representation.

1 INTRODUÇÃO

1.1 Ontologias

De origem filosófica, o estudo das ontologias busca compreender a realidade dos seres, sua essência, princípios e causas (SEARLE, 1998), baseando-se em exaustivas classificações dos elementos do mundo real visando identificar seu tipo (categoria), estrutura, propriedades, eventos, processos e as inter-relações estabelecidas entre estes elementos (SMITH, 2003).

Atualmente, estes estudos têm sido caracterizados pela interdisciplinaridade, envolvendo campos como a Ciência da Informação (VICKERY, 1997; SOERGEL, 1999; ALMEIDA e BAX, 2003; ALMEIDA, 2013) e a Ciência da Computação (WAND e WEBER, 1990; GUARINO, 1998; WAND, STOREY e WEBER,1999; CANTELE *et al*, 2004; FONSECA, 2007), com implicações em domínios da medicina (FARINELLI *et al*, 2016; MENDONÇA e ALMEIDA, 2016, ALMEIDA e FARINELLI, 2017), gestão e organização do conhecimento (TORRES, SIMÕES e ALMEIDA, 2017), biologia (HARRIS e PARKINSON, 2003), engenharia, gestão pública (FARINELLI, SILVA e ALMEIDA, 2013), direito (BREUKERS e HOEKSTRA, 2004; GRIFFO, ALMEIDA e GUIZZARDI, 2015) e geografia (RODRIGUES, CHAVES e SILVA, 2006; MACHADO *et al*, 2011).

Em função do desenvolvimento da Web Semântica (BERNERS-LEE; HENDLER e LASSILA, 2001) e da necessidade de lidar com a recuperação adequada de massivas informações digitais (ALMEIDA; SOUZA e BARACHO, 2015), as ontologias têm sido utilizadas como recurso alternativo para a representação da realidade em função da formalidade com que seus modelos são concebidos, baseados em lógica descritiva e axiomas formais. Quando convertidos em artefatos de informação proveem vocabulários semânticos consensuais que viabilizam a interoperabilidade semântica entre sistemas computacionais.

Na área da saúde, conforme pesquisas realizadas nas últimas décadas, grande esforço tem sido destinado a integrar e recuperar dados para uso profícuo destes sistemas, especialmente visando a criação do prontuário eletrônico de pacientes - EHR (RECTOR, 2000; SMITH *et al*, 2007; ANDRADE, ALMEIDA e SCHULZ, 2012; FARINELLI et al, 2016; MARTÍNEZ-COSTA, C., SCHULZ, 2017; BODENREIDER, 2018; HAENDEL *et al*, 2018).

Em função da complexidade dos dados e informações biomédicas (insumos para a construção de ontologias), compreende-se a relevância das pesquisas no entorno de seu ciclo de vida. Embora a incipiência de tais pesquisas, iniciativas de aspectos complementares podem ser observadas.

Em Ibraim e Ataelfadiel (2017), e Noy *et al* (2010), o ciclo de vida é visto sob o enfoque principal da Engenharia de Ontologias, subárea da Engenharia do Conhecimento e que aborda processos, métodos e metodologias de desenvolvimento e disseminação de ontologias, bem como plataformas tecnológicas relacionadas (NOY *et al*, 2010). Embora em Noy *et al* (2010) as análises ocorram sobre a construção de ontologias biomédicas, apenas em Ibraim e Ataelfadiel (2017) encontram-se aportes teóricos para classificação dos dados em um ciclo de vida de ontologias.

Em VUMC (2005) encontra-se o modelo "*Translational Research*" que visa a descrição do processo de incorporação de resultados de uma pesquisa científica às práticas reais e, a partir delas, a retroalimentação de novas pesquisas. O modelo compreende aspectos relacionados aos processos de acesso a dados, sendo composto por quatro ciclos sequenciais que abrangem a coleta, o consumo e a geração de insumos para um novo processo ou pesquisa.

Em Hunkler *et al* (2011), encontra-se a apresentação de um processo de governança de grandes volumes de dados provenientes da área da saúde, coletados via *surveys* (*Survey of Health, Ageing and Retirement in Europe* - SHARE). O processo envolve as tarefas e atores relacionados à governança, além de cuidados especiais para a disseminação pública de dados sensíveis.

Na seção seguinte serão apresentados aspectos teóricos acerca dos dados e ciclos de vida que serviram de subsídio para análise de aderência das etapas de construção de uma ontologia biomédica e de identificação dos aspectos de qualidade resultantes.

1.2 Dados e Ciclos de Vida

O conceito de dado tem sua referência histórica na teologia, sua utilização tanto no singular, quanto no plural é encontrada em registros que datam do século XVII (BORGMAN, 2015). É importante ressaltar que em sua percepção, o significado do dado não reside nele mesmo, sendo reflexo de um contexto, um símbolo, um referencial que depende da interpretação do observador.

Embora pareça uma abstração óbvia em nosso dia a dia, o dado não possui uma definição simples, devido sua ampla abrangência semântica. Ele é o insumo base tanto para os sistemas de informação, quanto para os instrumentos informacionais como Listas Controladas, Taxonomias, Tesauros e Ontologias. Constitui-se de uma unidade celular que acrescida de metainformação e semântica, contribui para que a informação seja produzida. Esta visão estruturada em camadas (dado - informação - conhecimento) possui ampla aceitação nos estudos e práticas referentes ao estudo da informação e do conhecimento.

No entanto, sua definição tem sido simplificada e associada a exemplos como números, símbolos, letras e fatos (*National Research Council* 1999, apud Borgman). O delineamento de suas fronteiras conceituais muitas vezes não são claras e se mesclam com conceituações sobre informação. Neste trabalho um conceito norteador para dado é o utilizado pela *Organisation for Economic co-operation and Development* (OECD[1]) em sua definição sobre "dado de pesquisa"; a organização integrada por 36 nações membro, expressa que: dados de pesquisa são definidos como registros factuais (pontuações numéricas, registros textuais, imagens e sons) usados como fontes primárias para pesquisa científica, e que são comumente aceitos na comunidade científica como necessários para validar os resultados de uma pesquisa. Um conjunto de dados de pesquisa constitui uma representação parcial e sistemática do assunto investigado.

O dado, em comparação com outras estruturas que lidam com informação, como sistemas computacionais por exemplo, também possui "Ciclo de Vida": Um modelo sequencial de ações que resultam em decisões na utilização/manutenção do dado. Briney (2015) elenca uma

1 https://www.oecd.org/sti/sci-tech/38500813.pdf

comparação entre ciclo de vida antigo e novo na perspectiva do dado. Na Figura 1, do Apêndice A, são representadas as quatro etapas consideradas como ciclo de vida antigo, sendo que na primeira etapa, "Planejamento do Projeto", é necessário pensar em quais dados devem ser adquiridos. O próximo passo é trabalhar a "Aquisição dos Dados" propriamente dita. Em seguida é vista a "Análise dos Dados" para que a última etapa, "Publicação" possa ser realizada. A autora destaca que neste modelo composto por quatro estágios, a aquisição e análise são críticos para o sucesso do processo, e sua crítica reside no sentido de que este ciclo não recompensa o uso do dado além da análise, considerando-os como um subproduto da pesquisa ao invés de serem produtos com importância em si mesmos, como um artigo por exemplo. Este modelo foi usado por centenas de anos, mas com o advento digital, o ciclo demanda outras etapas, abordadas no Novo Ciclo de Vida.

Este novo conjunto de processos considera "O Compartilhamento de Dados", a "Preservação" e o "Reuso" como passos no processo de pesquisa, além de incluir "Planejamento de Gerenciamento de Dados e Projeto", "Aquisição de Dados", "Análise de Dados", "Publicação de Artigos" e "Divulgação de Dados", "Preservação" e "Reuso". Na Figura 2, do Apêndice A, são apresentados estes processos e seu fluxo. Este ciclo assemelha-se a outras propostas, como a da DataONE (2018), demonstrada na Figura 3 no mesmo apêndice.

Para Mearian (2003), o ciclo de vida informacional também deve abordar sua persistência, o suporte para sua gravação e recuperação. Observou-se nos artigos pesquisados que a preocupação com a questão processual das etapas de tratamento do dado em seu ciclo de vida foi uma etapa posterior, anteriormente havia uma atenção em relação à arquitetura de armazenamento. Período aquele em que muitas informações foram perdidas por falta de *hardware* compatível com mídias substituídas por outras. Políticas de migração de dados entre sistemas e sua possível interoperabilidade entre aplicações diversas eram importantes para minimizar o impacto da evolução tecnológica.

Diversos autores e instituições propuseram ciclos de vida de dados abordando processos que, em várias etapas se assemelham. Santana (2013) destaca que não é possível abordar um processo de ciclo de vida sem considerar o contexto em que os dados devam ser tratados, bem como sua especificidade, seja administrativa, descritiva, técnica, estrutural ou de preservação. O autor ainda propõe um Ciclo de Vida dos Dados para Ciência da Informação, CVD-CI, representado na Figura 4, do Apêndice A.

Em Lenhardt *et al* (2014) encontram-se passos comuns em várias propostas de ciclo de vidas observadas e já descritas anteriormente, tais como: Planejar, Coletar, Controlar Qualidade, Documentar, Preservar, Usar.

Justifica-se o investimento na adoção de um ciclo de vida de dados por benefícios como: facilidade na recuperação, uso e análise dos dados, tanto pelo pesquisador como por colaboradores; fornecimento de dados de qualidade para pesquisas translacionais; melhoria na tomada de decisão; incremento do valor investido na pesquisa; expansão da compreensão sobre um determinado problema e/ou domínio, etc. Um gerenciamento eficiente pode permitir que cientistas externos ao projeto possam encontrar, entender e utilizar os dados no futuro. Adicionando a este cenário a documentação por metadados, a possibilidade de citações e créditos futuros é garantida, além de permitir a rastreabilidade autoral.

2 OBJETIVOS

Embora robustas as pesquisas voltadas à aplicação de ontologias na área da saúde, as respectivas práticas de governança sobre os dados tratados nestas pesquisas precisam ser aprimoradas. Conforme apresentado anteriormente, não foram identificadas práticas de governança do ciclo de vida do dado dentro do ciclo de vida ontológico, sendo este o fator motivador da presente pesquisa.

Desta forma, o objetivo deste estudo será o de compreender o ciclo de vida de dados biomédicos a fim de representá-los em uma ontologia e, por extensão, compreender ainda, de que forma a gestão deste ciclo de vida poderá agregar maior qualidade à representação de tais dados.

Para o estudo foi selecionada a Ontologia Obstétrica e Neonatal - OntoNEO[2] (FARINELLI, ALMEIDA, ELKIN e SMITH, 2016), uma ontologia realista que abrange os domínios de cuidados no pré-natal, intraparto e pós-natal. Tal escopo foi estabelecido baseando-se na análise de um conjunto de padrões de EHR (nacionais e internacionais), sistemas de EHR adotados por diferentes hospitais e entrevistas com obstetras americanos e brasileiros visando identificar os fluxos de trabalho das clínicas de saúde e as demandas informacionais dos médicos envolvidos. A ontologia foi construída por membros do grupo de Pesquisa RECOL[3] da Universidade Federal de Minas Gerais (UFMG), do qual os autores do presente artigo participam.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa de natureza qualitativa e caráter descritivo baseou-se em uma análise dos procedimentos que conduziram a construção da OntoNEO, em comparação com as etapas dos ciclos de vida previamente apresentados neste estudo. Estes procedimentos foram organizados em etapas selecionadas dos estudos de Briney (2015), DataONE (2018) e Dias (2018). Por fim foram analisados aspectos de qualidade agregados à OntoNEO em função das práticas adotadas em sua construção. Os resultados serão apresentados na próxima seção.

4 RESULTADOS

Os resultados das análises realizadas serão apresentados em cinco etapas, a saber: 1) Informações sobre dados e formatos de dados, 2) Conteúdo dos metadados e formatos, 3) Políticas para acesso, compartilhamento e reuso, 4) armazenamento de longo prazo e, 5) questões orçamentárias.

2 https://bioportal.bioontology.org/ontologies/ONTONEO
3 http://recol.eci.ufmg.br/

4.1 Informações sobre dados e formatos de dados

A pesquisa foi realizada a partir do consumo e da produção de dados experimentais variados, como:

i. Dados brutos ou derivados:

a) consumo de dados provenientes dos padrões adotados para a definição de seu escopo e para prover insumos de dados para a construção da ontologia: Registro de Saúde da Mulher, Registro Anteparto e Formulário Pós-Parto fornecidos pelo Colégio Americano de Obstetras e Ginecologistas (ACOG); o Formato de Registro Eletrônico de Saúde das Crianças fornecido pela Agência de Pesquisa e Qualidade em Assistência à Saúde (AHRQ); Padrões para a Estrutura Clínica e Conteúdo dos Registros de Pacientes fornecidos pelo Centro de Informação de Saúde e Assistência Social do Reino Unido e pela Academia de Faculdades Reais de Medicina; Padrão Nacional para Informações de Resumo de Alta do Paciente da Autoridade Australiana de Informações e Qualidade em Saúde. Observou-se a preocupação com as especificidades terminológicas, técnicas e administrativas do contexto, o que, em termos de qualidade, contribuiu para a expansão da compreensão sobre o domínio do problema e para a padronização dos insumos de dados para a construção da ontologia. Tais práticas se assemelham aos processos de Planejamento e Coleta encontrados em DataONE (2018);

b) consumo de dados provenientes de pesquisa translacional: um conjunto de 11 ontologias abertas, previamente construídas e publicadas no Bioportal[4] serviram de base para a construção da OntoNEO. Tais artefatos proveram classes, entidades e relacionamentos para reuso, em aderência aos princípios da OBO Foundry[5] para a construção de ontologias interoperáveis voltadas à representação de realidades biológicas e biomédicas. Entre elas encontram-se a Ontologia de Artefatos e Informação (IAO), a Ontologia dos Genes (GO) e a Ontologia de Doenças Humanas (DOID). Neste sentido observou-se a preocupação com a qualidade dos dados brutos a serem reutilizados, com a perspectiva de prover, ao término da construção, uma ontologia de qualidade para alimentar novas pesquisas translacionais, conforme visto em VUMC (2005).

ii. Modelos e saídas

a) Produção de modelo de representação do conhecimento abrangendo coleções físicas de classes, atributos e relacionamentos, distribuídas em três conjuntos de dados: informações básicas sobre os cuidados de saúde pertencentes ao escopo da ontologia; dados de formulários e documentos relacionados e entidades sociais envolvidas no contexto. A OntoNEO encontra-se disponível publicamente no repositório de ontologias biomédicas Bioportal. Neste aspecto observou-se a preocupação com as especificidades técnicas e administrativas para que a ontologia pudesse ser publicada respeitando-se as normas de qualidade da OBO Foundry. Tal cuidado constituiu em um

4 https://bioportal.bioontology.org/
5 http://www.obofoundry.org/

incremento do valor investido na pesquisa uma vez que possibilita à comunidade externa o acesso e reúso de sua produção. As práticas adotadas assemelham-se ao processo de Publicação/Divulgação de Dados descrito no novo modelo de Briney (2015);

b) Realização de Prova de Conceito (PoC) sobre a capacidade da ontologia em prover interoperabilidade semântica no domínio estabelecido. Estes resultados podem ser encontrados em Farinelli, Almeida e Emygdio (2018), tendo por saída a proposição de uma arquitetura de interoperabilidade semântica. Observou-se a preocupação com a validação do dado produzido (OntoNEO) quanto à proposta de se constituir em um artefato de informação realmente capaz de integrar semanticamente sistemas computacionais. A PoC garante a eficácia da ontologia produzida e a qualidade do processo de construção.

iii. Materiais científicos

a) Publicação de artigos científicos relacionados à pesquisa: no período de 2013 a 2017, relativo à produção da pesquisa de doutorado que culminou na construção da OntoNEO, foram produzidos 11 artigos, direcionados a periódicos científicos e congressos e simpósios de grande reconhecimento na área. Dentre eles destaca-se a publicação no JASIST (ALMEIDA e FARINELLI, 2017). Neste aspecto observou-se a preocupação quanto a divulgação dos conhecimentos teóricos e empíricos que poderão nortear a condução de novas pesquisas, na comunidade nacional e internacional. Tais atividades relacionam-se às previstas em Briney (2005), e compreendem uma contrapartida relevante às agências financiadoras, além de agregarem valor ao reconhecimento da pesquisadora..

iv. Sobre processamento dos dados

A construção da OntoNEO baseou-se em um modelo de ciclo de vida iterativo e incremental, provendo desenvolvimento gradativo da ontologia, com a definição de novas entidades e relações a cada iteração. Três níveis conceituais foram desenvolvidos antes de sua formalização: a) definição do escopo da iteração; b) avaliação de aderência às ontologias preexistentes no domínio para a definição de vocabulário consensual Os seguintes critérios foram estabelecidos para a seleção de classes para reuso: i) a ontologia em questão é aderente aos princípios da OBO Foundry?; ii) as classes possuem definição?; iii) quantas vezes o termo foi previamente reutilizado? iv) qual o status da ontologia no repositório da OBO Foundry?; c) construção de novos termos e relações para atender aos requisitos da OntoNEO não cobertos na fase anterior. As ferramentas OntoFox e Protegé foram utilizadas para importação dos termos e axiomas necessários para a geração do arquivo em formato OWL da OntoNEO. O controle de versões da ontologia foi realizado a partir do ambiente colaborativo para construção de ontologias CMAP Tools.[6] Cada versão foi submetida para validação por representantes de diferentes comunidades de especialistas, incluindo médicos. Há compromisso do autor principal em manter atualizações desta ontologia, em função dos avanços científicos na área. O controle de qualidade realizado nestes procedimentos repousam sobre os dados de reuso para construção da OntoNEO, integração de versões, e preservação da ontologia, consonância com Briney (2005), DataONE (2018) e Dias (2018).

6 https://cmap.ihmc.us/

4.2 Conteúdo dos metadados e formatos

Conforme citado anteriormente, foram utilizados diversos padrões para formalização de conteúdos de metadados e formatos, além dos princípios da OBO Foundry, essenciais para a construção de ontologias realistas de alto nível agregáveis ao repositório em questão. Houve preocupação em relação à qualidade de representação dos dados, aumentando a possibilidade de reuso futuro e realimentação de pesquisas translacionais e independentes de domínio em função do nível da ontologia.

4.3 Políticas para acesso, compartilhamento e reuso

A OntoNEO segue ainda as diretrizes da OBO Foundry para acesso, compartilhamento e reuso. Os direitos autorais são garantidos aos autores e devem ser mantidos pela licença Creative Commons (**License:**CC-BY). Tais práticas, associadas às mencionadas no item 4.2 contribuem para facilitar a rastreabilidade autoral, garantir maior número de citações e créditos futuros. E ainda, aproximam-se dos processos de controle encontrados em Dias (2018).

4.4 Armazenamento de longo prazo

A ontoNEO é um projeto que se encontra na fase inicial de desenvolvimento, podendo ser acessada pelos sites: https://ontoneo.com/ e https://bioportal.bioontology.org/ontologies/ONTONEO, onde será mantida em caráter permanente. Os aspectos de qualidade envolvidos neste item relacionam-se aos processos de preservação do dado, políticas de arquivamento, padrões da comunidade e atribuição de responsabilidades, encontrados em Briney (2005) e Dias (2018).

4.5 Questões orçamentárias

O projeto foi mantido em parte por meio de financiamento das agências de fomento CAPES e CNPQ ao autor principal. Outra parte do projeto foi financiada pelo NIH NCATS sob o prêmio CTSA Número UL1TR001412. As questões orçamentárias são previstas no plano de gestão de dados proposto por Dias (2018).

5 CONSIDERAÇÕES FINAIS

O presente artigo apresentou uma revisão teórica acerca dos estudos sobre ontologias e sobre os ciclos de dados de pesquisa buscando traçar um caminho de compreensão da relevância da

adoção das práticas de governança de dados como premissa para a atribuição de maior qualidade aos processos de construção de ontologias biomédicas.

Uma análise descritiva foi realizada sobre todas as fases de construção da ontologia OntoNEO, abrangendo desde a aquisição de variados dados brutos, puros ou provenientes de pesquisas translacionais, até a publicação das diversas saídas do projeto. Estas fases, apresentadas na seção de resultados caracterizam um tipo de ciclo de vida dos dados para a construção de ontologias biomédicas de maior qualidade, conforme discutido em cada análise.

Em função do volume e do valor resultados oriundos da pesquisa, especialmente por sua aceitação como uma ontologia de alto nível, integrante do projeto OBO Foundry, é possível concluir que a adoção das práticas de governança, incluindo as de versionamento e atualização das ontologias constituem uma iniciativa relevante para agregar maior qualidade a estes artefatos, tornando mais eficaz sua capacidade de prover interoperabilidade semântica entre sistemas computacionais.

REFERÊNCIAS

ALMEIDA, M. B. (2013). Revisiting Ontologies: a necessary clarification. **Journal of the American Society of Information Science and Technology**. v. 64, n. 8. p. 1682–1693. DOI: 10.1002/asi.22861.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ci. Inf.**,Brasília, v. 32, n. 3, p. 7-20, set./dez. 2003.

ALMEIDA, M. B.; FARINELLI, F. (2017). Ontologies for Representation of Electronic Medical Records: the Obstetric and Neonatal Ontology. **Journal of the Association for Information Science and Technology**. doi:10.1002/asi.23900.

ALMEIDA, M. B.; SOUZA, R.; BARACHO, R. M. A. Looking for the identity of Information Science in the age of big data, computing clouds and social networks. In: **Proceedings of the 4th International Symposium of Information Science (ISI 2015)**, Croatia.

ANDRADE, A.Q.; ALMEIDA, M.B.; SCHULZ, S. Revisiting ontological foundations of the OpenEHR Entry Model. **Proceedings of the International Conference of Biomedical Ontologies (ICBO)**, 2012, Austria.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**., v.1, p. 34 - 43, 2001.

Best Practices | DataONE. Disponível em: <https://www.dataone.org/best-practices>. Acesso em: 26 set. 2018.

BORGMAN, C. L. **Big data, little data, no data: scholarship in the networked world**. The MIT Press: England, 2015.

BREUKERS, J. A. P. J.; HOEKSTRA, R. J. Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law. **CEUR Workshop Proceedings**, 2004.

BRINEY, K. **Data management for researchers: organize, maintain and share your data for research success**. Exeter, UK: Pelagic Publishing, 2015.

CANTELE, R. C.; ADAMATTI, D. F.; FERREIRA, M. A. G. V.; SICHMAN, J. S. Reengenharia e ontologias: análise e aplicação. **I Workshop de Rede Semântica (WWS2004)**, 1, 2004.

DIAS, G. A. **Guia de Dados de Pesquisa**, 2018. Acesso em: 22 ago. 2018.

FARINELLI, F.; ALMEIDA, M. B.; ELKIN, P.; SMITH, B. OntONeo: The Obstetric and Neonatal Ontology. Proceedings of the **International Conference of Biomedical Ontologies**, Corvallis, U.S, 2016.

FARINELLI, F.; ALMEIDA, M.B.; ELKIN, P.; SMITH, B. (2016). Dealing with Social and Legal Entities in the Obstetric and Neonatal Domain. In Proceedings of the International Conference of Biomedical Ontologies, Corvallis, U.S.

FARINELLI, F.; ALMEIDA, M.B. EMYDGIO, J.L. Arquitetura de dados orientada por ontologias: interoperando sistemas de saúde. In: XVI Congresso Brasileiro de Informática em Saúde (CBIS 2018).

FARINELLI, F.; SILVA, S. de M.; ALMEIDA, M. B. Iniciativas governamentais para interoperabilidade semântica. **Revista Fonte**, ano 10, v. 13, p.66-74, dez. 2013.

FONSECA, F. The double role of ontologies in information science research. **Journal of the American Society for Information Science and Technology**, v. 58, p. 786-793, 2007.

GRIFFO, C.; ALMEIDA J. P. A.; GUIZZARDI, G.. Towards a legal core ontology based on alexy's theory of fundamental rights. In: **15th International Conference on Artificial Intelligence and Law (ICAIL)**, 2015.

GUARINO, N. Formal ontology and information systems. In N. Guarino (Ed.). **Formal ontology in Information Systems** (1998, pp. 3–15). Amsterdam: IOS Press.

HAENDEL, M. A.; MCMURRY, J. A.; RELEVO, R.; MUNGALL, C. J.; ROBINSON, P. N.; CHUTE, C. G.. A Census of Disease Ontologies. **Annual Review of Biomedical Data Science**, v.1, , p. 305-331, 2018.

HARRIS, M.; PARKINSON, H.. Standards and Ontologies for Functional Genomics:Towards Unified Ontologies for Biology and Biomedicine. **Comp. Funct. Genom.** 2003; 4:116–120.

IBRAHIM, E. A. A.; ATAELFADIEL, M. A. M. Ontology Life Cycle: A Survey on the Ontology and its Development Steps. International Journal of Science and Research (IJSR), v. 6, n. 8, 2017, p. 1352-1357.

LENHARDT, W. C.; AHALT, S.; BLANTON, B.; CHRISTOPHERSON, L.; IDASZAK, R. Data Management Lifecycle and Software Lifecycle Management in the Context of Conducting Science. **Journal of Open Research Software**, v. 2, n. 1, p. e15, 9 jul. 2014.

MACHADO, I. M. R.; ALENCAR, R. O. de; J.; CAMPOS, R. de O.; JUNIOR, C.I A. D.. An ontological gazetteer and its application for place name disambiguation in text. **Journal of the Brazilian Computer Society**, 17(4):267-279, 2011.

MARTÍNEZ-COSTA, C.; SCHULZ, S. Validating EHR clinical models using ontology patterns. **Journal of Biomedical Informatics**, 76 (2017) 124–137.

MEARIAN, L. The Data Life Cycle. **Computerworld**, Knowledge Center Storage. v. November 17, 2003, p. 39, 40, 2003.

NOY, N.; TUDORACHE, T.; NYULAS, C.; MUSEN, M. The ontology life cycle: Integrated tools for editing, publishing, peer review, and evolution of ontologies. **AMIA Annual Symposium Proceedings**, v. 2010, p. 552–556, 2010.

RECTOR, A. L. **Clinical terminology: why is it so hard?** Methods Inf Med. 1999 Dec;38(4-5):239-52.

RODRIGUES, C.; CHAVES, M. S.; SILVA, M. J. Uma Representação Ontológica da Geografia Física de Portugal. **IX Encontro de Utilizadores de Informação Geográfica**, ESIG-Oeiras, Portugal, 2006.

SANTANA, R. C. G. Ciclo de Vida dos Dados e o Papel da Ciência da Informação. **XIV Encontro Nacional de Pesquisa em Ciência da Informação (Enancib 2013)**, p. 21, [s.d.], 2013.

SMITH, B.; ASHBURNER M.; ROSSE, C.; BARD, J.; BUG, W.; CEUSTERS, W.; GOLDBERG, L. J.; EILBECK, K.; IRELAND, A.; MUNGALL, C. J.; The OBI Consortium; LEONTIS, N.; ROCCA--SERRA, P.; RUTTENBERG, A.; SANSONE, S.; SCHEUERMANN, R. H.; SHAH, N.; WHETZEL, P.; LEWIS, S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. **Nat Biotechnol**. 2007 November; 25(11): 1251. doi:10.1038/nbt1346.

SOERGUEL, D. **The rise of ontologies or the reinvention of classification**.1999. Disponível em: http://www.dsoergel.com/cv/B70.pdf. Acesso em: out. 2018.

STRASSER, C.; COOK, R.; MICHENER, W.; BUDDEN, A. Primer on Data Management: What you always wanted to know. **DataONE**, p. 11, [s.d.].

TORRES, S.; SIMÕES, M.G.; ALMEIDA, M.B. (2017). Princípios para modelagem de domínios em sistemas de organização do conhecimento In: **Anais do III Congresso ISKO Espanha-Portugal**.

VICKERY, B. C. Ontologies. **Journal of Information Science**, London, v. 23, n. 4, p. 227-286, 1997.

VUMC - Vanderbilt University Medical Center **Strategic Plan for VUMC Informatics.** 2005. Disponível em: <https://ncs.mc.vanderbilt.edu/Data/NonSecure/IC_Strategic_Plan_9-12-05.pdf>.

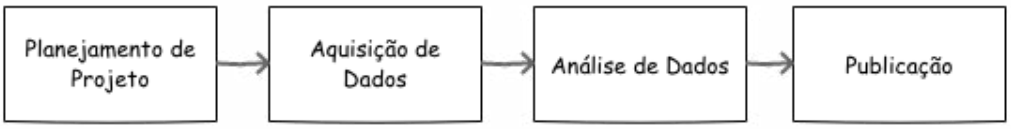
WAND, Y.; WEBER, R. An Ontological Model of an Information System. **IEEE Transactions on Software Engineering**, Vol. 16, No. 11, November 1990.

_____. An Ontological Analysis of the Relationship Construct in Conceptual Modeling. **ACM Transactions on Database Systems**, Vol. 24, No. 4, December 1999, Pages 494–528.

WAND, Y.; STOREY, V. C.; WEBER, R. An Ontological Analysis of the Relationship Construct in Conceptual Modeling. **ACM Transactions on Database Systems**, Vol. 24, No. 4, December 1999.

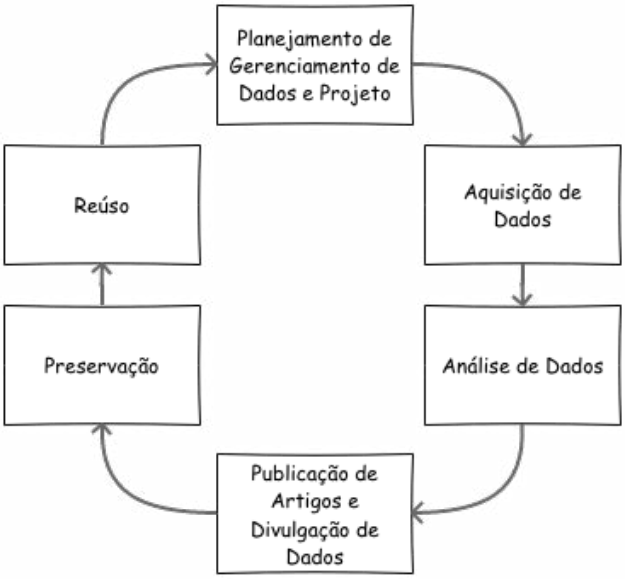
APÊNDICE A

Figura 1 - Ciclo de vida antigo



Fonte: Adaptado de Briney (2015)

Figura 2 - Ciclo de vida novo

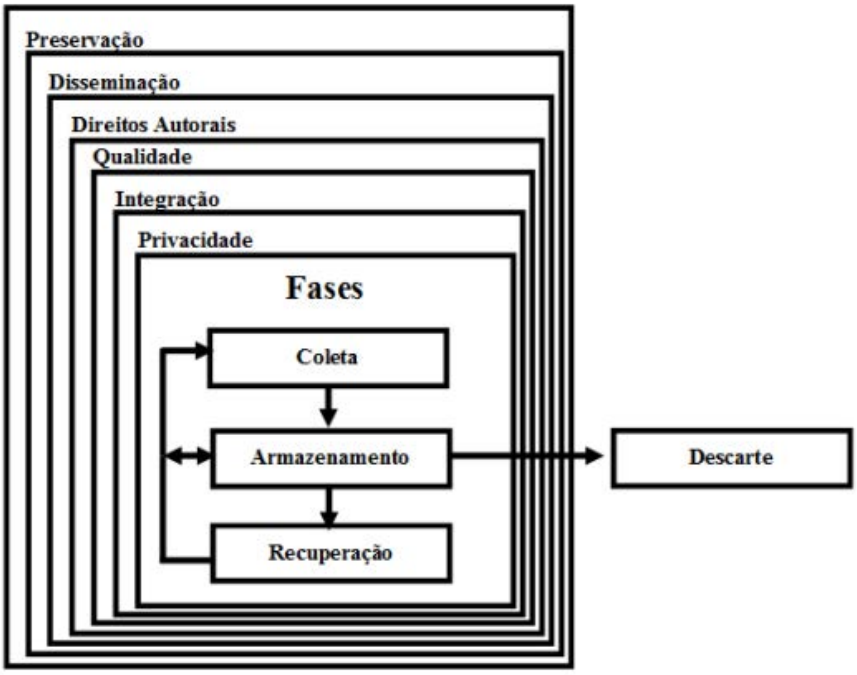


Fonte: Adaptado de Briney (2015)

Figura 3 - Ciclo de vida dos dados

Fonte: DataONE (2018)

Figura 4 - Ciclo de Vida dos Dados para Ciência da Informação, CVD-CI



Fonte: Santana (2013)

CIÊNCIA ABERTA:
os desafios para concepção da ciência cidadã

OPEN SCIENCE:
the challenges for the conception of citizen science

Luciana Gonçalves Silva Souza[1], Izabel Antonina de Araújo[2], Júlia Rocha Ribeiro[3]
(1) Doutoranda do Programa de Pós-Graduação em Gestão & Organização do Conhecimento pela Universidade Federal de Minas Gerais. E-mail: lucianags9@gmail.com.
(2) Doutoranda do Programa de Pós-Graduação em Ciência da Informação pela Universidade Federal de Minas Gerais. E-mail: antoninaizabel@gmail.com.
(3) Mestranda do Programa de Pós-Graduação em Gestão & Organização do Conhecimento pela Universidade Federal de Minas Gerais. E-mail: julia34rr@gmail.com.

Resumo:
Nas últimas décadas, a ciência aberta é tema discutido nas agendas dos governos e organizações mundiais. Esta ciência envolve vários aspectos do processo de pesquisa que se baseia nas novas formas de produção de pesquisa colaborativa, interativa e compartilhada com o cidadão. Nesta perspectiva, países como Nova Zelândia, Reino Unido, Colômbia, Japão, Brasil entre outros, estão desenvolvendo vários projetos que promovem o engajamento dos cidadãos na ciência como alternativa para o desenvolvimento de uma cultura de conscientização científica. Esta participação nas decisões e nos processos de seleção dos fluxos de pesquisa que se dá não apenas como observadores e coletores de dados, mas também como profissionais planejadores e avaliadores. Para isso, necessitam de informações de qualidade, plataformas interativas, gestão de dados, entre outros. Nesse sentido, o objetivo deste trabalho pontuar aspectos relacionados aos pilares que compõem a ciência aberta e identificar as ações desenvolvidas por alguns países que visam envolver o cidadão no fazer científico. Trata-se de uma pesquisa qualitativa, de caráter exploratório a partir de uma revisão bibliográfica sobre o tema e identificação de projetos de ciência cidadã praticados em alguns países. Observa a importância da participação cidadã na definição das práticas a serem desenvolvidas no âmbito da ciência, tecnologia e inovação. Considera que pensar em uma ciência aberta requer outras dimensões e condições consideradas essenciais para uma ciência cidadã, como acesso aberto, plataformas abertas, agenda de pesquisa abertas, além da informação de qualidade e literacia científica.
Palavras-chave: Ciência aberta. Ciência cidadã. Acesso aberto. Dados abertos. Participação cidadã.

Abstract:
In the last decades, open science is a topic discussed in the agendas of governments and organizations worldwide. This science involves several aspects of the research process that is based on the new forms of production of collaborative, interactive and shared research with the citizen. In this perspective, countries such as New Zealand, the United Kingdom, Colombia, Japan, Brazil among others, are developing several projects that promote citizen engagement in science as an alternative to the development of a culture of scientific awareness. This participation in the decisions and in the process of selection of the research flows that occurs not only as observers and collectors of data, but also as professional planners and evaluators. For this, they need quality information, interactive platforms, data management, among others. In this sense, the objective of this work is to highlight aspects related to the pillars that compose open science and to identify the actions developed by some countries that aim to involve the citizen in the scientific making. This is a qualitative research, exploratory in nature, based on a bibliographical review on the subject. It observes the importance of citizen participation in defining the practices to be developed in the scope of science, technology and innovation. It considers that thinking about an open science requires other dimensions and conditions considered essential for a citizen science, such as open access, open platforms, open research agenda, and quality information and scientific literacy.
Keywords: Open Science. Citizen science. Open access. Open data. Citizen participation.

1 INTRODUÇÃO

A pesquisa científica produz diversos dados e instrumentos que subsidiam o seu desenvolvimento. Com o avanço da ciência e das tecnologias de informação e comunicação, a produção de grandes volumes de dados passa a se configurar como uma preocupação com relação à organização, guarda e difusão das informações para a sociedade.

O movimento de disponibilização dos produtos utilizados e resultantes das investigações científicas passa a ser ressaltado com o conceito de ciência aberta, que pode ser entendida a partir dos seguintes pilares:

- Acesso aberto;
- Dados abertos;
- Investigação / Inovação aberta;
- Redes abertas de ciência;
- Ciência cidadã (PORTUGAL, c2016b).

Dessa forma, constitui a abertura de todo o processo científico, ressaltando, também, o compromisso da ciência com a sociedade. A implantação da ciência aberta impulsiona a inovação e desenvolvimento de produtos, serviços e oportunidades de negócios (PORTUGAL, c2016a). Compreender esses diferentes propósitos da ciência aberta constitui importante caminho para o desenvolvimento de práticas mais inclusivas para compreensão e participação da sociedade nos processos científicos.

2 OBJETIVOS

Este trabalho tem como objetivo pontuar aspectos relacionados aos pilares que compõem a ciência aberta, enfatizando o desenvolvimento da ciência cidadã, por razão de sua importância para a aproximação entre comunidade científica e sociedade, buscando fazer com que seja compreensível a importância de ambos no desenvolvimento das ações cotidianas.

3 PROCEDIMENTOS METODOLÓGICOS

O presente artigo foi desenvolvido utilizando como método a pesquisa qualitativa, de caráter exploratório, pois o tema ainda é pouco explorado no cenário nacional. Foi realizada uma revisão bibliográfica tendo por referência artigos científicos e documentos governamentais orientadores, buscando compreender a ciência aberta e ciência cidadã e a aplicação prática desse conceito, bem como a identificação de alguns projetos de ciência cidadã praticados por alguns países

4 RESULTADOS

Apresenta-se, nessa sessão, as informações obtidas a partir do levantamento bibliográfico proposto.

4.1 Ciência aberta

A ciência aberta propõe a abertura e a divulgação das atividades vinculadas ao campo científico, fazendo com que todos os atores envolvidos nas investigações compreendam com clareza e acompanhem todo o processo realizado.

As ações de ciência aberta precisam incluir aspectos relacionados aos denominados pilares de sua sustentação: acesso aberto, dados abertos, investigação / inovação aberta, redes abertas de ciência e ciência cidadã, conceitos que serão brevemente apresentados a seguir.

O acesso aberto diz respeito à livre disponibilização de documentos científicos na internet, seja por meio de sua difusão em plataformas digitais, como os repositórios, que podem ser institucionais ou temáticos (chamada via verde, conforme a Budapest Open Access Iniciative), ou pela publicação dos conteúdos em periódicos de acesso aberto (via dourada) (MARRA, 2014).

Essas estratégias de acesso aberto aumentam a visibilidade, o acesso e o uso das produções científicas, favorecendo o progresso da ciência e o acompanhamento das atividades realizadas, ao oferecer à sociedade acesso facilitado aos documentos produzidos.

O processo de abertura de dados parte de um movimento ocasionado pela identificação da necessidade de disponibilizar conteúdos produzidos por segmentos sociais como agências governamentais, órgãos de pesquisa e indústrias, constituindo informações fundamentais para o desenvolvimento científico (SANTOS, 2017).

A investigação / inovação aberta compreende aspectos relacionados à docência e às práticas educativas voltadas à melhoria do ensino e aprendizagem, promovendo retorno científico a partir da compreensão e aplicação de diferentes práticas com resultados positivos desenvolvidas em diferentes países, contextos e formas de ensino (POMBO, 2018).

As redes abertas de ciências constituem grupos de investigadores em colaboração que compartilham e constroem novos conhecimentos. São desenvolvidas pesquisas e projetos que buscam integrar especialidades e conhecimentos diferentes em prol de um objetivo (PEREIRA, 2018).

A ciência cidadã, assunto que permeia as discussões do presente artigo, compreendendo a participação da sociedade na construção de conteúdos científicos, será detalhado na próxima sessão, com o objetivo de definir e ilustrar o conceito com algumas iniciativas mundiais.

4.2 Ciência Cidadã

A ciência cidadã está relacionada ao envolvimento do cidadão nos processos científicos, abrange uma grande variedade de atividades, algumas mais ligadas à educação, outras à prática, e alguns que misturam ambos - como educação, aprendizagem, ensino e prática em ciências. Esse envolvimento se estrutura de forma dinâmica em diferentes aspectos de engajamento que tem sido identificado com relação ao grau de participação do público e o papel dos cidadãos na tomada de decisão e processo de seleção dos fluxos de pesquisa que receberão financiamento público.

Para Tiago (2012), o envolvimento de não cientistas na investigação científica teve um grande crescimento e desenvolvimento nos últimos anos. Para a autora, algumas questões científicas só podem ser respondidas com a recolha e tratamento de uma grande quantidade de dados, mas,

devido a uma escassez de recursos humanos e financeiros, a contribuição dos cidadãos na realização de tarefas relacionadas aos projetos de pesquisa auxilia o trabalho dos pesquisadores.

Conforme Maida (2007), nos últimos anos a revolução digital está mudando o estilo de vida da sociedade, principalmente no que se refere a capacidades de cidadania e progresso científico. A exploração de recursos naturais aumentou a necessidade de um desenvolvimento sustentável que atenda às necessidades dos usuários sem comprometer a capacidade de atender as necessidades das gerações futuras. Goodchild (2007) *apud* Tiago explana que os projetos de pesquisa utilizam os cidadãos como "sensores" que contêm informações locais ricas e com potencial para fornecer informações em grandes áreas ou para ser atualizada em tempo hábil.

Sobre este aspecto, Albagli (2015), observa a diversidade que caracteriza a ciência cidadã em seus amplos aspectos e significados que podem ser compreendidos em duas vertentes: "uma delas, conhecida como *crowdsourcing*, reúne iniciativas que buscam mobilizar contribuições voluntárias, de vários tipos, em esforços de pesquisa, por parte de não cientistas, incluindo desde compartilhamento de recursos computacionais até coleta de informações de relevância científica" (p.15). Na outra vertente, conforme elucida Albagli, estão iniciativas orientadas para "maior participação, intervenção e empoderamento de cidadãos não só nas formas de produção e uso, mas nos próprios rumos da pesquisa" (p. 15).

Sob o entendimento de que o interesse e a participação do público nas ações científicas aumenta a confiança nas pesquisas e contribui para o avanço científico, muitos países vêm desenvolvendo projetos científicos em que o cidadão tem um papel importante, conforme descrito no tópico seguinte.

4.2.1 Exemplos de projetos de desenvolvimento da ciência cidadã

Alguns exemplos de grandes projetos de ciência cidadã executados a nível mundial são relatados no documento Open research agenda setting, da Organisation for Economic Co-operation and Development (OECD), publicado em dezembro de 2017. A organização realizou seis estudos de caso, que são relatados abaixo.

Caso 1 - Projeto CIMULACT, Citizen and Multi-Actor Consultation on Horizon 2020 - projeto realizado em 30 países europeus, envolvendo mais de 1000 cidadãos, buscando estabelecer o entendimento sobre políticas, pesquisa e inovação, possibilitando diálogos que promovem a pesquisa científica. Para capacitar os cidadãos foram realizados workshops e conferências, com recriação de cenários de pesquisa, fazendo com que os cidadãos compreendessem como poderiam contribuir com suas visões, necessidades e preocupações (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2017).

Caso 2 - Great New Zealand Science Project (GNZSP). Desenvolvido na Nova Zelândia, o projeto teve como objetivo incentivar a participação pública nas pesquisas, por meio de desafios científicos nacionais em diversas áreas, como biodiversidade, alterações climáticas, alimentos, saúde. A atividade proposta pelo Ministério dedicado às questões de inovação e emprego do país foi veiculada na mídia, com a participação de um jovem e um pesquisador, que indicavam um site que apontava os desafios e propunha ao cidadão a construção de um projeto científico próprio. Ao final do projeto, todas as sugestões foram analisadas e avaliadas pelo setor de ciência e inovação. A iniciativa teve por objetivo possibilitar a socialização de

ideias e a compreensão dos desafios da ciência, permitindo ao governo compreender as necessidades da sociedade (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2017).

Caso 3 - Ideas for Change (IFC). O projeto, desenvolvido na Colômbia, buscou incentivar a participação de comunidades vulneráveis no desenvolvimento de práticas científicas. Os cidadãos foram convidados a contribuir especificando suas necessidades e fornecendo desafios concretos, além de participando das etapas de seleção e desenvolvimento de soluções inovadores em ciência e tecnologia. Buscou integrar as comunidades vulneráveis e os membros da comunidade científica, contando com o apoio técnico e financeiro de órgãos governamentais e organizações do país. As iniciativas possibilitaram a confiança, coesão social e inclusão no contexto da ciência, tecnologia e inovação (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2017).

Caso 4 - James Lind Alliance (JLA). O projeto foi iniciado em 2004 pelo National Institute for Health Research Evaluation da University of Southampton no Reino Unido, mas teve suas atividades expandidas para o Canadá, Holanda e Alemanha. As atividades têm como objetivo identificar e priorizar questões levantadas por pacientes, médicos e demais profissionais da saúde com relação aos tratamentos médicos e doenças. As principais perguntas orientam workshops com participação de pacientes e profissionais de saúde. Com a realização das atividades é possível identificar questões que necessitam de investimento em pesquisa, conforme evidências e dúvidas que precisam ser priorizadas (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2017).

Caso 5 - RISTEX. A RISTEX é uma agência japonesa de financiamento de projetos em pesquisa e desenvolvimento que incentiva soluções inovadoras de enfrentamento de problemas sociais. A agência realiza, a cada dois anos, a seleção cuidadosa de questões sociais que necessitam de investimento. As propostas devem se enquadrar em critérios como envolvimento de todas as partes interessadas na implementação das ações, adoção de abordagens multidisciplinares em pesquisa e inovação, aplicação em várias regiões e comunidades, ampla divulgação dos resultados. O objetivo é estreitar a relação entre pesquisadores e cidadãos, mas segundo análise da OECD esse impacto ainda é limitado, devido à necessidade de adaptação da metodologia e ampliação dos incentivos para participação dos pesquisadores (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2017).

Caso 6 - X-Project. O projeto, desenvolvido na Coreia, parte da apresentação de problemas de pesquisa pelo cidadão, que são resolvidos cientificamente pelos pesquisadores. Busca favorecer o envolvimento de toda a sociedade na identificação de problemas que necessitam de solução científica, ampliando a comunidade de pesquisa. O projeto apresenta princípios que devem ser seguidos, como o respeito e recompensa a todas as ideias, a busca de solução a todos os problemas, a abertura do conteúdo do projeto ao público, a simplicidade na estrutura e metodologia de solução. O objetivo do projeto é buscar o interesse e a participação do público nas ações científicas, aumentando a confiança nas pesquisas realizadas e o entendimento das práticas científicas (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2017).

Como iniciativas brasileiras relacionadas à ciência cidadã elencam-se as definidas no domínio do Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr), serviço do Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC). Esse sistema agrega uma infraestrutura de dados e conteúdos em biodiversidade e busca integrar e capacitar interessados, incluindo segmentos de pesquisadores, estudantes, educadores, sociedade civil, órgãos governamentais, não governamentais e empresas do setor privado. O SiBBr apresenta vários projetos em biodiversidade, dados ecológicos,

catálogo de espécies, registro de coleções brasileiras. São listados 16 projetos brasileiros relativos à ciência cidadã, os quais estão sintetizados no Quadro 1, abaixo.

Quadro 1 - Projetos brasileiros de desenvolvimento da ciência cidadã

Projeto	**Objetivo**
Observatório de Imprensa, Avistamento e Ataques (OIAA Onça)	Ampliar o conhecimento sobre carnívoros silvestres brasileiros.
Portal de Zoologia de Pernambuco	Compartilhamento de informações sobre animais selvagens, realizando um inventário da biodiversidade de forma lúdica e divertida.
EXOSS	Registro de meteoros na atmosfera, utilizando uma rede de voluntários, para a pesquisa de chuvas de meteoros e identificação de seus corpos parentais.
Sistema Urubu	Avaliar os efeitos de infraestrutura viária na biodiversidade mundial. Nesta fase o principal foco é o atropelamento de animais selvagens.
Táxeus	Promover conhecimento da biodiversidade brasileira a partir de listas de espécies.
WIKIAVES	Apoiar, divulgar e promover a atividade de observação de aves, fornecendo gratuitamente ferramentas avançadas para controle de fotos, sons, textos, identificação de espécies.
Atlas de Registros de Aves Brasileira	Atlas reúne informações referentes à biologia das espécies e ações de conservação.
Eu vi uma ave usando pulseira?!	Compreender qual é o tamanho da área de vida de cada ave na paisagem agrícola por meio da ação de observadores de aves brasileiros.
Cidadão Cientista	Realizar monitoramentos participativos de aves em Unidades de Conservação e parques urbanos.
AETRAPP	Engajar comunidades no monitoramento de populações de mosquitos Aedes, vetores de zika, dengue, chikungunya e febre amarela urbana, por meio de uma armadilha caseira, um aplicativo e um mapa online.
Guardiões da Chapada	Monitorar os visitantes florais e flora associada na Chapada Diamantina, sensibilizar o público acerca da importância do serviço de polinização e comunicar temas relacionados à ecologia e conservação.
Brydes do Brasil	Elaborar uma base de dados sobre as baleias-de-bryde identificadas individualmente a partir de fotos da espécie em águas brasileiras.
Blue Change	Conservar ambientes marinhos e costeiros com o auxílio de pesquisadores, instituições e organizações sociais. Conecta pessoas com a ciência, levando demandas da sociedade para que os pesquisadores as transformem em objeto de investigação.
Onde estão as baleias e os golfinhos?	Incentivo à postagem de fotos e/ou vídeos de cetáceos (baleias, botos e golfinhos) em águas costeiras do Rio de Janeiro.
ECOA	Auxiliar comunidades tradicionais e assentados da reforma agrária na identificação de áreas com potencial de reflorestamento e/ou restauração e no mapeamento de árvores frutíferas nativas de relevância ambiental e econômica para a coleta de frutos e sementes.
Ciência Cidadã - UFABC	Implantar o programa de educação científica em diferentes contextos com a criação, aplicação e avaliação de protocolos de ciência cidadã visando promover a alfabetização científica de diferentes públicos, trabalhar a sua percepção ambiental e possibilitar a coleta de dados sobre a biodiversidade.

Fonte: BRASIL, [2018].

Assim, as ações de ciência cidadã desenvolvidas buscam estabelecer, de formas diferentes, a integração entre a comunidade científica e o cidadão, fazendo com que esse compreenda as ações que são realizadas pelos pesquisadores e a importância de sua participação na definição das práticas a serem desenvolvidas no âmbito da ciência e tecnologia.

5 CONSIDERAÇÕES FINAIS

A ciência aberta está sendo discutida dentro de um processo de mudança social em que busca-se fazer a interlocução da ciência com a sociedade considerando os diversos tipos de saberes para o enfrentamento dos problemas sociais que se tornaram desafios globais da humanidade.

A evolução da ciência cidadã é evidenciada pelos projetos apresentados neste trabalho, os quais elucidam ações estratégicas realizadas em vários países na tentativa de envolver os cidadãos nas questões científicas. Essa participação se dá por meio de observação, coleta, registros, análise de dados em diversas áreas do conhecimento. Essa socialização de ideias, além de aproximar o cidadão do fazer científico, promove a visibilidade de grupos afetados, diminui os custos das pesquisas e emerge uma cultura de compartilhamento.

A contribuição do cidadão é primordial para evolução coletiva dos problemas e formulação de hipóteses, uma vez que a grande quantidade de dados impossibilita que as equipes de pesquisa consigam explorar todas as possibilidades de investigação.

Dessa forma, observa-se que o cidadão pode contribuir muito para o avanço científico, mas para essa contribuição ser igualitária e eficiente é necessário pensar em outras dimensões consideradas bases para a ciência cidadã como acesso aberto, plataformas abertas, agenda de pesquisa abertas, além da informação de qualidade e literacia científica.

REFERÊNCIAS

ALBAGLI, Sarita. *Ciência aberta, questões abertas. In:* ALBAGLI, Sarita; MACIEL, Maria Lucia; ABDO, Alexandre Hannud (Org.). **Ciência aberta, questões abertas**. Brasília: IBICT; Rio de Janeiro: UNIRIO, 2015. 312 p.

MAIDA, C. A. (Ed.). **Sustainability and communities of place**. New York and Oxford: Berghahn. 2007.

MARRA, Patrícia dos Santos Caldas. Visibilidade dos repositórios institucionais brasileiros: análise de diretórios internacionais de acesso aberto. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, v. 8, n. 3, mar. 2014. Disponível em: <https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/672>. Acesso em: 03 out.2018.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). **Open research agenda setting**. 21 Dec. 2017. Disponível em: <http://www.oecd.org/about/> Acesso em: 23 ago. 2018.

PEREIRA, Luísa Álvares. O desenvolvimento da escrita ao longo do currículo: do ensino básico ao ensino superior. In: **II FÓRUM CIDTFF**: construindo um compromisso com a ciência aberta: livro de posters. Universidade de Aveiro, 2018. p. 36. Disponível em: <https://ria.ua.pt/bitstream/10773/23715/3/documento.pdf>. Acesso em: 05 out. 2018.

POMBO, Lúcia et al. Práticas de ciência aberta no EduPark. In: **II FÓRUM CIDTFF**: construindo um compromisso com a ciência aberta: livro de posters. Universidade de Aveiro, 2018. p. 22. Disponível em: <https://ria.ua.pt/bitstream/10773/23715/3/documento.pdf>. Acesso em: 05 out. 2018.

PORTUGAL. Ministério da Ciência, Tecnologia e Ensino Superior. Fundação para a Ciência e a Tecnologia. **Glossário**. Lisboa, c2016a. Disponível em: <http://www.ciencia-aberta.pt/glossario>. Acesso em: 22 ago. 2018.

PORTUGAL. Ministério da Ciência, Tecnologia e Ensino Superior. Fundação para a Ciência e a Tecnologia. **Sobre ciência aberta**. Lisboa, c2016b. Disponível em: <http://www.ciencia-aberta.pt/sobre-ciencia-aberta>. Acesso em: 22 ago. 2018.

TIAGO, Patrícia Maria Nunes. **Improving citizen science as a tool for biodiversity monitoring**. Tese (Doutorado em Biologia) – Faculdade de ciências, Universidade de Lisboa, 2012. Disponível em: <http://repositorio.ul.pt/bitstream/10451/34139/1/ulsd731841_td_Patricia_Tiago.pdf>.

SANTOS, Paula Xavier (Coord.). **Livro Verde Ciência aberta e dados abertos**: mapeamento e análise de políticas, infraestruturas e estratégias em perspectiva nacional e internacional. Rio de Janeiro: Fiocruz, 2017. 141 p.

BRASIL. Ministério da Ciência, Tecnologia, Inovações e Comunicações. Sistemas de Informação sobre a Biodiversidade Brasileira. **Ciência cidadã**: projetos. [2018]. Disponível em: <http://www.sibbr.gov.br/cienciacidada/>. Acesso em: 12 nov. 2018.

DESAFIOS TECNOLÓGICOS NO PROCESSO DE RECEPÇÃO NORMATIVA: adequação e conformidade na perspectiva da Lei de Proteção de Dados Pessoais

TECHNOLOGICAL CHALLENGES IN THE NORMATIVE RECEPTION PROCESS: adequacy and compliance from the perspective of the Law of Protection of Personal Data

Andrea Carmo Name Willemin[1] Geralda Magella de Faria[2] Cláudio José Amante[3]
(1) Universidade Federal de Santa Catarina (UFSC), Programa de Pós-Graduação em Ciência da Informação, Centro de Ciências da Educação, Bloco B, sala 105, Campus Universitário, Trindade, Florianópolis, andrea.willemin@gmail.com
(2) Universidade Federal de Santa Catarina (UFSC), Programa de Pós-Graduação em Direito, Centro de Ciências Jurídica (CCJ), Campus Universitário, Trindade, Florianópolis, geraldamagella@gmail.com
(3) Universidade Federal de Santa Catarina (UFSC), Coordenador do Programa de Pós-Graduação em Administração Universitária, Professor do Departamento de Odontologia, Campus Administração Universitária, Trindade, Florianópolis, claudio020461@gmail.com

Resumo: (1) A Lei de Proteção de Dados, recém editada no Brasil, tem a marca de resguardar, proteger e cuidar dos dados – entendidos estes pertinentes à inovação e tecnologia, de cunho sensível, privado e pertinentes à pessoa - com a medida do zelo e da segurança que até então não foram reservados pela legislação. (2) A pesquisa levada a termo pretende ser exploratória, qualitativa e que terá como base uma proposta de resolução de problemas. (3) Neste sentido, a adequação tecnológica e organizacional requer a indicação de um "modelo de processo de resolução de problemas", cujo fundamento deve prestar ao regramento normativo, cumplicidade e conformidade até então inexistente. (4) Assim, processo na esfera de proteção de dados requer procedimentos que busquem inovação e configuração tecnológica apta ao desempenho seguro de dados pessoais.
Palavras-chave: Lei de Proteção de Dados Pessoais. Privacidade; Inovação. Política Nacional de Ciência, Tecnologia e Inovação.

Abstract: (1) The Data Protection Law, recently published in Brazil, has the mark of safeguarding, protecting and taking care of the data - understood as pertinent to innovation and technology, sensitive, private and pertinent to the person - with the measure of zeal and security that were not previously reserved by law. (2) The research carried out to term is intended to be exploratory, qualitative and based on a proposal for problem resolution. (3) In this sense, the technological and organizational adequacy requires the indication of a "problem resolution process model", the foundation of which should be to regulate law, complicity and compliance that did not exist until then. (4) Indeed, process in the sphere of data protection requires procedures that seek innovation and technological configuration for the safe performance of personal data.
Keywords: Law of Protection of Personal Data. Privacy. Innovation. National Policy on Science, Technology and Innovation.

1 INTRODUÇÃO E METODOLOGIA

Com os dados ganhando destaque e espaço na cena internacional e brasileira, saber interpretar e aplicar as diferentes leis que dispõe sobre Proteção de Dados Pessoais é essencial para atuar neste novo cenário regido pela confluência do universo jurídico e tecnológico. A GDPR, Lei Europeia de Proteção de Dados Pessoais, e a brasileira, no caso a Lei de Proteção de Dados Pessoais LPDP - Lei 13709, de 14 de agosto de 2018, possuem inúmeros desafios de caráter tecnológico e jurídico.

Vários pontos da lei chamam particularmente a atenção em razão do caráter inovador e desafiador que trarão quanto a sua integração nos diferentes sistemas de informação e organizacionais gerados pelas novas exigências legais. Várias nações regularam a proteção de dados pessoais há décadas atrás. Hoje, por exemplo, a Europa possui uma legislação forte que impõe diferentes metodologias nas quais a proteção de dados pessoais é pensada desde a concepção de sistemas, práticas comerciais, projetos, produtos ou qualquer outra solução que envolva o manuseio de dados pessoais.

O Brasil aprovou recentemente sua Lei de Proteção de Dados Pessoais. Trabalhar com os dados pessoais agora exige especial desafio para que a proteção dispensada pela lei seja realizada sem que com isso se perca o caráter inovador e realizador buscado pelas novas tecnologias em desenvolvimento. Por inovação, entende-se a atividade de caráter científico, tecnológico, organizacional, financeiro ou comercial proposta e executada com o objetivo de obter produtos, processo tecnológico e serviços totalmente novos ou melhorados, que pode se dar em regime de transação comercial ou não. Inovação é definida como resultado da introdução de novo conhecimento ou tecnologia econômica e socialmente útil (o sentido do novo se aplica ao local onde é introduzido e não possui sentido universal). (BRASIL, 2018).

Para tanto é necessário criar processos que permitam as organizações integrarem dentro de suas estruturas as novas regras legais. Os desafios são inúmeros e em diferentes aspectos. A figura do apêndice A demonstra, de forma exemplificava, as três dimensões (organizacionais, tecnológicas e humanas) necessárias de figurarem de forma harmônica no sistema de informação possibilitando assim as soluções de negócios desejadas. Neste estudo serão levantados alguns dos desafios de dimensões tecnológicas impostos pela lei geral de proteção de dados pessoais que precisarão ser implantados nas diferentes organizações a partir de janeiro de 2020. A complexidade do desafio trazido pela LPDP exige num primeiro ponto que se identifique o que é necessário ser feito. Pois são muitas as novidades introduzidas pela lei e que passarão a ser cobradas. Assim sendo convém a apresentação às organizações de um panorama geral do que está sendo solicitado em termos de adaptações tecnológicas para que daí se estabeleçam processos de adequação das diferentes implementações exigidas pela lei.

No tocante aos procedimentos metodológicos do presente estudo, trata-se de uma pesquisa exploratória, que aborda certos desafios tecnológicos encontrados quando da integração com os novos dispositivos legais trazidos pela lei de proteção de dados pessoais. Da mesma forma pode ser entendida como uma pesquisa qualitativa e de natureza básica que terá como base o passo 01 (um) "modelo do processo de resolução de problemas" (LAUDON; LAUDON, 2014, p. 20).

2 OBJETIVOS

O objetivo geral deste estudo é identificar os desafios tecnológicos no processo de recepção normativa da Lei de Proteção de Dados Pessoais relacionando aos principais desafios de dimensões tecnológicas decorrentes da Lei 13709/2018, para a criação de um "modelo de processo de resolução de problemas" (LAUDON; LAUDON, 2014, p. 20).

Este estudo é a primeira etapa necessária para se "entender e solucionar problemas organizacionais por meio dos sistemas de informação" (Apêndice B), que poderá ser aplicado dentro das diferentes organizações que terão o desafio, até janeiro de 2020, de implantarem a nova Lei de Proteção de Dados Pessoais brasileira. Para que se possa resolver problemas, neste caso, particularmente os referentes as exigências legais, é necessário, primeiramente, que exista consenso sobre sua existência, suas causas e o que pode ser feito quanto a eles, levando-se em conta os recursos limitados que as organizações dispõe (LAUDON; LAUDON, 2014, p. 20) e a necessidade das organizações se adaptarem. A identificação de alguns dos desafios tecnológicos da Lei de Proteção de Dados Pessoais nos sistemas de informação poderá, com o auxílio da ferramenta do modelo de "processo de resolução de problemas" (Apêndice C) num primeiro plano, auxiliar a evidenciar os pontos que necessitam serem trabalhados, para que se possa, depois estabelecer diferentes propostas; avaliar e escolher as propostas e implantar as melhores soluções nas organizações públicas e privadas de forma a se cumprir as novas exigências tecnológicas impostas pela lei.

3 CIÊNCIA DA INFORMAÇÃO E PROTEÇÃO DE DADOS

Foi uma longa gestação, da pedra ao byte, e da descoberta e reconhecimento quanto a importância da seleção, arquivamento, transmissão e o próprio aprendizado, entre prática e teoria até dar o passo do reconhecimento da Ciência da Informação.

Na história humana, tudo começou com a oralidade, depois vieram os registros. Rudimentares aos olhos atuais, mas de significado incomensurável para a humanidade. Cita-se, o Cilindro de Ciro "o Grande", primeiro rei da Pérsia, que em 539 a.c conquistou a Babilonia, e recomendou o vetusto registro que é reconhecido como a 1ª. Carta de Direitos Humanos (BRASIL, 2018). Também a Pedra da Roseta, decifrada recentemente, entre 1802 a 1822, por Chanpollion, e sua inscrição registra um decreto do rei Ptolomeu V, promulgado em 196 a.C., na cidade Mênfis. É certo que o desejo político, no caso, a necessidade do registro e a transmissão às gerações fizeram de "dados" como este, e os milhares que se seguiram, uma qualidade e condição de seu armazenamento e transmissão, quer pelo trabalho, como também, pela própria necessidade humana de organização.

Os primeiros cientistas da informação, conforme revela Araújo (2018), tiveram reconhecimento somente no Século XX, de forma que de 1920 a 1940, da Inglaterra até aos EUA e em outros países, os cientistas puderam abastecer seus colegas - com os Science Services - em suas respectivas áreas de estudos, com índices, resumos, canais, arquivos. Assim, os "ex" químicos, físicos, engenheiros e outros cientistas alçaram a condição de cientistas da Informação, saíram da prática para o universo da institucionalização.

De forma concreta, em 1968, o American Documentation Institute, dos Estados Unidos, mudou seu nome para American Society for Information Science, tornando-se a primeira instituição de Ciência da Informação do mundo. Contribuíram também para sua formação, primeiro a Inglaterra, com a realização da Royal Society Scientific Information Conference, em 1948, e a criação, em 1958, do Institute of Information Scientist. Pouco depois, na União Soviética, foi criado o Viniti, Vserossiisky Institut Nauchnoi i Tekhnicheskoi Informatsii, vinculado à Academia de Ciências. E, a seguir, em 1958, ocorreu nos Estados Unidos a International Conference on Scientific Information, quando, a Ciência da Informação alça a condição de ciência dedicada à informação em ciência e tecnologia, com a preocupação de base dessas ações: não mais a necessidade de se ter a posse dos documentos, mas a prioridade dada à sua circulação, ao seu fluxo, e ao atendimento das necessidades dos cientistas, inclusive com a vertente póscustodial de Documentação, até que, em 1962 foi publicado um estudo, de autoria de Machlup, voltado a produção e distribuição de conhecimentos. Em 1963, um relatório de Weinberg apontou que as agências de governo fomentadoras de pesquisa científicas deveriam também assumir a responsabilidade pela transferência do conhecimento gerada nas pesquisas. (ARAÚJO, 2018).

Aqui tem-se a ponta de lança, na medida em que vários pesquisadores passam a estudar o processo da "comunicação da informação científica": o estudo dos vários registros do cientista, da ideia na sua cabeça (relatórios, seminários, apresentações em eventos, artigos em periódicos, livros, citações ao trabalho, menções em livros-textos e enciclopédia), até as suas características, vantagens e desvantagens, tempo médio de produção, entre outros aspectos, inclusive, amparar e buscar os dados, estejam estes em bibliotecas, em arquivos, em museus, em bases de dados, em artigos de periódicos, onde estiver, um estudioso da ciência da informação estará de "olhos abertos" e dedos disponíveis para cataloga-los. Qual seja, a "comunicação informal" passa a consolidar a concepção da Ciência da Informação e seu objeto de estudo - dos fluxos, dos caminhos percorridos pela informação, sua materialização em diferentes produtos e serviços – passa a ser uma importante lição e fundamentação à legislação, da qual a nova LPDP, é, sem dúvida, uma destacada resposta.

4 LEI DE PROTEÇÃO DE DADOS PESSOAIS: considerações e desafios

Devem ser destacados na LPDP alguns aspectos vinculados às tecnologias e que são por ela normatizados.

Estes pontos foram identificados como desafios tecnológicos a serem suplantados em razão da necessidade das novas exigências legais, de forma a se integrarem ou coexistirem com os sistemas de informação atuais e, neste aspecto, anunciar o estado da arte.

4.1 Anonimização (art. 11, § 3º, da Lei 13709/2018).

O processo de eliminação ou modificação da informação pessoal existente numa base de dados, com o objetivo de dificultar ou impedir a identificação unívoca dos indivíduos.

A referida lei dispõe sobre padrões e técnicas utilizados em processos de anonimização e regulamenta as possíveis verificações acerca de sua segurança. O grande desafio tecnológico

ainda é como utilizar os dados mesmo que anonimizados, por exemplo, no processo de busca de determinados dados sem comprometer a anonimização

De acordo com a utilização de técnicas e o alcance dos objetivos, podem ser considerados três divisões: de-identificação, anonimização e pseudonimização.

A de-identificação é o processo de remoção ou ofuscação de toda a informação pessoal de uma base de dados, com o objetivo de impedir a identificação dos indivíduos. É possível reverter a de-identificação através de uma tabela de mapeamento (ligando os registros originais aos registros de-identificados). Ela suprime todos os atributos identificadores e também modifica os "*quasi-identifiers"*, mediante processos de generalização.

Um nível mais elevado da de-identificação é a anonimização, quando a pretensão é tornar impraticável, ou até mesmo impossível (com o uso de todos os meios considerados razoáveis) a re-identificação. Este processo impossibilitaria inclusive o próprio técnico que realizou a operação inicialmente de reverter o procedimento. Merece destaque o fato da definição ser adaptável ao contexto tecnológico do momento: "*todos os meios considerados razoáveis*". A afirmativa leva à ponderação de quais seriam os recursos necessários, o custo do processo e o conhecimento necessário para realizar uma re-identificação.

Já a pseudonimização objetiva substituir todos os identificadores pessoais por palavras ou códigos gerados artificialmente, os quais poderão funcionar como representações mascaradas dos dados originais. Uma pseudonimização relevante possui uma preocupação adicional relacionada com a incidência sobre os atributos "*quasi-identifiers*" (por exemplo a data de nascimento), e que a atribuição de códigos seja realizada de forma aleatória e independente dos valores originais (muito embora podendo ocasionalmente continuar relacionados entre si).

Infelizmente, diversos casos têm evidenciado as falhas em processos de anonimização. Estas situações têm evidenciada a necessidade do envolvimento de especialistas nos processos de anonimização, já que qualquer erro poderá gerar repercussões na imagem da organização.

A simples remoção de atributos identificadores (nome, telefone, etc.) nunca será suficiente para eliminar os riscos de quebra de privacidade. Por outro lado, é constante a possibilidade de um *hacker* acessar outras bases de dados externas, possibilitando o cruzamento dessa informação com os dados que se pretendam anonimizar (Pinho, 2017).

4.2 Direito ao esquecimento (art. 16, Lei 13709/2018).

O direito ao esquecimento diz respeito aos direitos privativos da personalidade, corolário do direito à intimidade, à privacidade, à honra e à imagem, e, na sociedade da informação, apesar de conservar a pertinência com o princípio da dignidade do ser humano, diz respeito às inúmeras notícias disseminadas pelos meios de comunicação e tem vínculo com a proteção de dados, pessoais e sensíveis, de onde decorrem inúmeros problemas de violação da personalidade acarretada pela perpétua exposição de dados.

Nesse contexto, os dados pessoais serão eliminados após o término de seu tratamento, no âmbito e nos limites técnicos das atividades, autorizada a conservação para algumas finalidades. Existem dificuldades técnicas para o seu cumprimento dada a grande quantidade de replicações e cópias de dados.

Os controladores são obrigados a fornecer relatórios de tratamento quando requisitados pelos titulares, sendo exigido não só uma plataforma de atendimento a essas requisições como também uma organização dessas informações internamente pelas instituições (art. 18).

Não só é necessário que a instituição consiga estruturar os dados requisitados de maneira que possa haver a portabilidade: ela deve prover os recursos técnicos e manter contato com outras instituições para que os canais de portabilidade funcionem corretamente (art. 18, inciso V).

4.3 Transferência internacional de dados (art. 33, Lei 13709/2018)

A transferência internacional de dados pessoais é permitida nos seguintes casos: a)para países ou organismos internacionais com grau de proteção de dados pessoais; b)quando o controlador apresentar garantias de cumprimento dos princípios, dos direitos e do regime de proteção de dados; c) quando a transferência for necessária para a cooperação jurídica internacional entre órgãos públicos de inteligência, de investigação e de persecução; d)quando a transferência for necessária para a proteção da vida ou da incolumidade física do titular ou de terceiro; e)quando a autoridade nacional autorizar a transferência e esta resultar em compromisso assumido em acordo de cooperação internacional; g)quando a transferência for necessária para a execução de política pública ou atribuição do serviço público; h)quando o titular tiver fornecido o seu consentimento específico para a transferência, com informação prévia sobre o caráter internacional da operação; i)quando necessário para atender as hipóteses seguintes: -para o cumprimento de obrigação legal ou regulatória pelo controlador; - para a execução de contrato ou de procedimentos preliminares relacionados a contrato do qual seja parte o titular, a pedido do titular dos dados; e -para o exercício regular de direitos em processo judicial, administrativo ou arbitral. Cumpre também destacar que, nos termos da Lei nº 12527/2011, as pessoas jurídicas de direito público, na esfera de suas competências e atividades, poderão requerer à autoridade nacional a avaliação do nível de proteção de dados pessoais dispensado pelo país ou organismo internacional.

No mesmo sentido da portabilidade, a transferência internacional de dados é um desafio para as instituições que necessitem/desejem efetuar a transferência não só pelos requisitos autorizados pela Lei quanto pelas dificuldades técnicas em efetuar essa empreitada.

4.4 Decisões automatizadas: (Art. 20, da Lei 13709/2018)

O titular dos dados tem direito a solicitar revisão, por pessoa natural, de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses, inclusive de decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade.

A possibilidade de rever decisões automatizadas será um grande desafio para as instituições que trabalham com significativos volumes de dados – em qualidade e quantidade - pois seus sistemas terão que conter informações suficientes que justifiquem a decisão para que a mesma possa ser revista, se for o caso. Em larga escala, o atendimento a essas demandas também terá que ser incorporado aos

sistemas utilizados. Esse parece ser um desafio enorme para instituições que não tem a preocupação de segurança em tecnologia de *Big Data*.

4.5 Relatórios de impacto (Art. 38, Lei13709/2018)

A autoridade nacional poderá determinar ao controlador que elabore relatório de impacto à proteção de dados pessoais, inclusive de dados sensíveis, referentes as operações de tratamento de dados, observados os segredos comercial e industrial das relações de negócios, e especialmente o sigilo e a confidencialidade entre pessoas e gestão.

As instituições deverão ter conhecimento e auditar seus sistemas para que os relatórios de impacto possam ser construídos, decorrentes de processos e procedimentos para o caso de vazamento de dados.

4.6 Boas práticas de segurança (Art. 46, Lei 13709/2018)

Os agentes de tratamento devem adotar medidas de segurança, técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou qualquer forma de tratamento inadequado com vistas a impedir a configuração de ilicitudes.

Para tanto, as atividades de tratamento de dados pessoais, inclusive as relativas aos dados sensíveis, devem primar pela observância do princípio da boa-fé, bem como, caracterizada, sua finalidade, adequação, necessidade, livre acesso, qualidade dos dados, transparência, segurança, prevenção, não discriminação; responsabilização e prestação de contas, no caso, a adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas.

É certo que pelo § 1º do art. 46 da LGPD, a autoridade nacional poderá dispor a respeito de padrões técnicos mínimos, visando tornar aplicáveis as disposições contidas no seu caput, para o qual levará em conta as seguintes situações: a natureza das informações tratadas, as características específicas do tratamento e o estado atual da tecnologia, especialmente.

Embora ainda não estejam regulados pela autoridade, as instituições deverão atender a padrões técnicos para o tratamento de dados pessoais. Pela lei, entende-se que a autoridade irá apontar regularmente as revisões em virtude da evolução da tecnologia e de novos padrões de segurança (Apêndice D).

4.7 Privacidade por *default* e por *design*: (Art. 46, § 2º Lei)

As medidas de que trata o caput deste artigo deverão ser observadas desde a fase de concepção do produto ou do serviço até a sua execução.

Tanto na concepção dos sistemas que fazem tratamento de dados quanto na sua execução, deverão ser observadas as garantias, prerrogativas e diretrizes da lei para o tratamento de dados pessoais. Isso significa uma revisão completa dos sistemas que fazem esse tipo de tratamento, assim como uma preocupação para os sistemas que estão em desenvolvimento e que virão a ser desenvolvidos. Em sistemas mais antigos e integrados em diferentes plataformas, isso pode significar uma revisão completa ou até mesmo uma necessidade de reestruturação de sistemas inteiros, em virtude da complexidade de se ter um sistema com proteção de privacidade desta forma.

4.8 Vazamento de dados: Art. 48.

O controlador deverá comunicar à autoridade nacional e ao titular a ocorrência de incidente de segurança que possa acarretar risco ou dano relevante aos titulares.

A construção dos relatórios de vazamento deverá ser pautada em informações acerca dos mesmos e nos riscos que eles representam, além das medidas que foram tomadas e serão tomadas acerca dos incidentes. Isso representa uma enorme mudança técnica no tratamento de dados pessoais das empresas, uma vez que a identificação de um vazamento, suas soluções técnicas imediatas e as mudanças que deverão decorrer em virtude dos mesmos representam um desafio tanto jurídico quanto tecnológico, pautando-se na segurança da informação.

5 RESULTADOS

Verifica-se que os objetivos propostos para o presente estudo foram atingidos uma vez que alguns dos principais problemas tecnológicos da lei foram elencados permitindo que se possa avançar, num futuro estudo, para a etapa dois, ou seja, para o "modelo de processo de resolução de problemas" (LAUDON; LAUDON, 2014, p. 20), pertinentes à implantação dos desafios tecnológicos nas organizações.

6 CONCLUSÃO

A identificação de alguns dos principais pontos dos desafios tecnológicos impostos pela Lei 13709/2018, permite a identificação de diferentes problemas que precisam ser resolvidos, e desta forma caracteriza a etapa 1 necessária para o "modelo processo de resolução de problemas" que poderá auxiliar as organizações a implantarem soluções adequadas e que satisfaçam as exigências legais de proteção de dados pessoais. Este primeiro passo é o ponto de partida para se criar um caminho rumo ao encontro das melhores soluções para a implantação da nova normatização de proteção de dados pessoais nas organizações.

REFERÊNCIAS

ARAÚJO, Carlos Alberto Ávila. **O que é Ciência da Informação?**. Disponível em http://www.uel.br/revistas/uel/index.php/informacao/article/view/15958/14205 Acesso em 15 nov 2018.

ASSOCIAÇÃO BRASILEIRA DE AGÊNCIAS DE PUBLICIDADE. **Cartilha Institucional de Segurança da Informação e Prevenção a Fraudes**. São Paulo. Disponível em < http://www.abap.com.br/pdfs/seguranca_informacao.pdf>. Acesso em 27/09/2018.

BRASIL. **Cilindro de Ciro**. Disponível em http://www.dhnet.org.br/direitos/anthist/marcos/cilindro/index.htm Acesso em 15 nov 2018.

BRASIL. **Lei n. 13709, de 14 de agosto de 2018**. Disponível em <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm>. Acesso em 27/09/2018.

BRASIL. Descritores em Ciências da Saúde. Disponível em http://decs.bvs.br/cgi-bin/wxis1660.exe/decsserver/ acesso em 08/10/2018.

BRASIL. Tribunal de Contas da União. Cartilha de boas práticas em segurança da informação - 4ª Edição. Brasília. Disponível em <https://portal.tcu.gov.br/biblioteca-digital/cartilha-de-boas-praticas-em-seguranca-da-informacao-4-edicao.htm>. Acesso em 07/10/2018.

CAVOUKIAN, Ann; JONAS, Jeff. **Privacy by design in the age of big data**. Ontário. Disponível em: < https://jeffjonas.typepad.com/Privacy-by-Design-in-the-Era-of-Big-Data.pdf>. Acesso em 07/10/2018.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de banco de dados**. Tradução de Daniel Vieira. São Paulo: Pearson, 2011.

HARBOUR, Pamela Jones. Privacy by design in law, policy and practice: a white paper for regulators, decision-makers and policy-makers. Ontário. Disponível em < http://www.ontla.on.ca/library/repository/mon/25008/312239.pdf>. Acesso em 07/10/2018.

LAUDON, Kenneth C.; LAUDON, Jane P. **Sistemas de Informação Gerenciais**. Tradução de Célia Taniwaki. São Paulo: Pearson Education do Brasil, 2014.

LOVELUCK, Benjamin. **Redes, liberdades e controle: uma genealogia política da internet**. Tradução de Guilherme João de Freitas Teixeira. Petrópolis: Vozes, 2018.

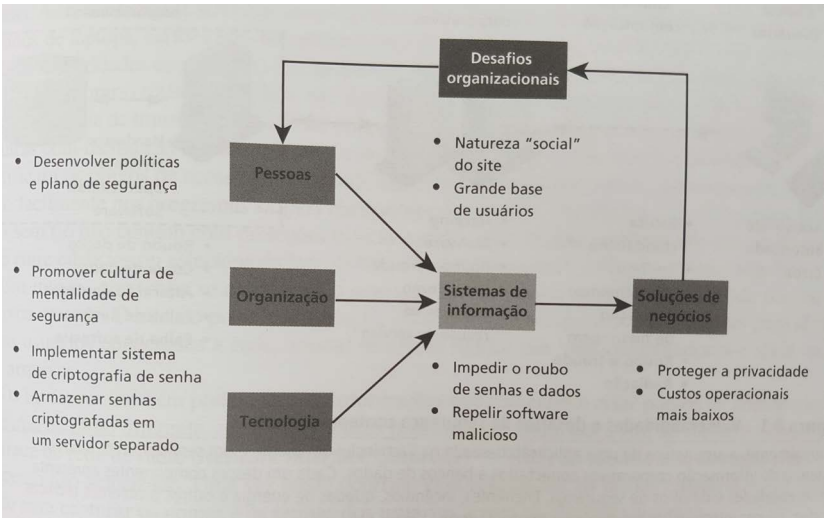
MARTINS, Claudia Aparecida et al. **Uma Experiência em Mineração de Textos Utilizando Clustering Probabilístico e Clustering Hierárquico.** São Carlos: Icmc, 2003.

PINHO, Frederico António Sá Oliveira. **Anonimização de bases de dados empresariais de acordo com a nova regulamentação europeia de proteção de dados**. Dissertação (Mestrado em Segurança Informática. Departamento de Ciência de Computadores. Faculdade de Ciências. Universidade do Porto. Porto, 2017.

PROVOST, Foster; FAWCETT, Tom. **Data Science para negócios**. Tradução de Marina Boscatto. Rio de Janeiro: Alta Books, 2016.

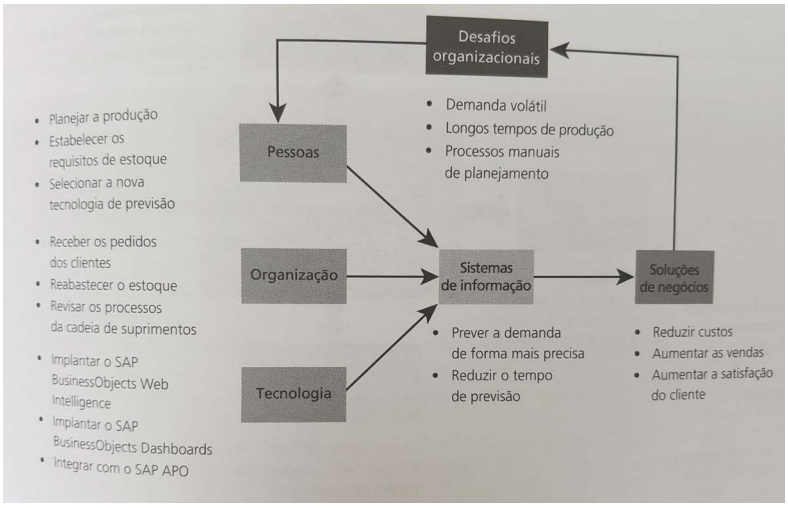
Sheshasaayee, Ananthi; Bhargavi. K. **A study of automated decision making systems**. Disponível em: < http://www.researchinventy.com/papers/v7i1/E07012831.pdf>. Acesso em 07/10/2018.
WIHLBORG, Elin; LARSSON, Hannu; HEDSTRÖM, Karin. "The Computer Says No!" - A Case Study on Automated Decision-Making in Public Authorities. Disponível em: <https://www.researchgate.net/publication/290303337_The_Computer_Says_No_--_A_Case_Study_on_Automated_Decision-Making_in_Public_Authorities>. Acesso em 07/10/2018.

Apêndice A –



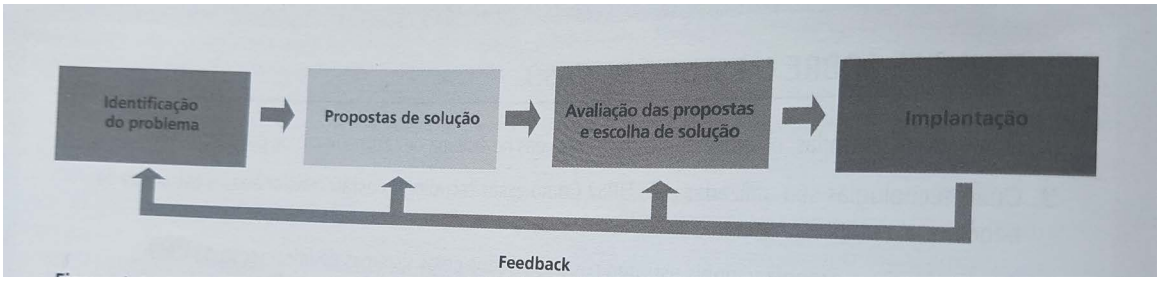
Sistemas de Informação Gerenciais, (LANDON; LANDON, 2014, p.255)

Apêndice B –



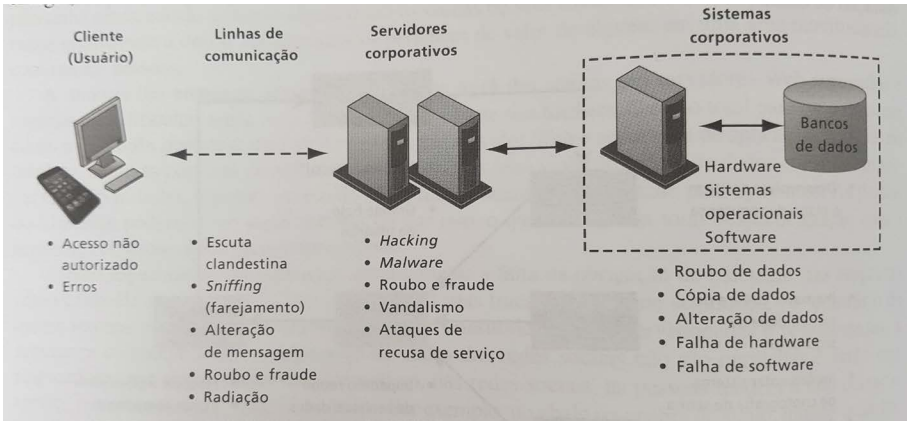
Sistemas de Informação Gerenciais, (LANDON; LANDON, 2014, p. 20)

Apêndice C –



Sistemas de Informação Gerenciais, (LANDON; LANDON, 2014, p. 20)

Apêndice D –



Sistemas de Informação Gerenciais (LANDON; LANDON, 2014, p. 256)

DESIGN AUDIOVISUAL PARA SISTEMAS GENERATIVOS:
Um experimento de construção audiovisual personalizada

AUDIOVISUAL DESIGN FOR GENERATIVE SYSTEMS:
A customized audiovisual construction experiment

Rafael M. Toscano[1], Valdecir Becker[1], Helder Bruno A. M. de Souza[1], Marianna C. Teixeira[1]
(1) Grupo de Pesquisa Design Audiovisual, Centro de Informática, Universidade Federal da Paraíba, João Pessoa - PB, Brasil, audiovisualdesign@lavid.ufpb.br

Resumo:
A produção de conteúdos audiovisuais tem se tornado uma atividade cada vez mais complexa a medida que novos recursos de interação, compartilhamento, diferentes telas e formas de visualização vem se popularizando. Diante deste cenário, surge a Arquitetura do Design Audiovisual (DA) como um instrumento teórico-metodológico de análise e criação de sistemas audiovisuais. O propósito deste artigo é descrever o processo de construção de um sistema audiovisual baseado em dados dos indivíduos para estabelecer uma fruição de conteúdo personalizada. A construção do sistema revelou que os elementos disponíveis no DA como linhas de design, papéis e gatilhos são elementos potenciais para criação deste tipo de sistema.
Palavras-chave: Design Audiovisual. Generativo. Personalização.

Abstract:
The production of audiovisual content has become an increasingly complex activity as new features of interaction, sharing, different screens and forms of visualization are becoming popular. In view of this scenario, the Architecture of Audiovisual Design (DA) arises as a theoretical-methodological instrument of analysis and creation of systems. The purpose of this article is to describe the process of building an audiovisual system that is based on individuals' data to establish a personalized content enjoyment. The construction of the system revealed that the elements available in the DA as design lines, roles and triggers are potential elements for the creation of this type of system.
Keywords: Audiovisual Design. Generative. Customization.

1 INTRODUÇÃO

O desenvolvimento de conteúdos audiovisuais tem se tornado uma atividade cada vez mais complexa a medida que novos recursos de interação, compartilhamento, diferentes telas e formas de visualização vem se popularizando. A experiência audiovisual, antes majoritariamente ocupada pelo cinema e televisão, hoje torna-se mais abrangente, podendo envolver desde pesquisas na Internet a complexos algoritmos de recomendação, colaboração e adaptação de conteúdo. Uma característica deste cenário é a presença e o uso de softwares para compor a experiência de fruição das obras audiovisuais. A expansão do poder de processamento e o aumento do acesso a uma conexão de rede estável viabilizam um cenário em que produtores de conteúdo audiovisual estão cada vez mais próximos de programadores, designers de interação e cientistas de dados.

Um vídeo online, filme, jogo ou mesmo uma novela, são exemplos de como as mídias de representação (áudio e vídeo) podem ser empregadas na construção de uma experiência estética e comunicacional. Se por um lado, temos as discussões acerca das tecnologias de captura e representação promovidas pela Computação, do outro temos as discussões no campo da Comunicação e Artes sobre os valores e significados que podem ser gerados pelo conteúdo que está sendo exposto.

A partir deste contexto computacional e midiático complexo é possível identificar discussões e experimentos que dialogam com a capacidade de geração de narrativas audiovisuais a partir da coleta de dados da audiência. Se nas primeiras produções cinematográficas e televisivas era inviável pensar a construção de conteúdos adaptados a interesses individuais ou de pequenos grupos, devido ao alto custo de produção e distribuição, hoje, com o aumento do processamento computacional e alcance de rede, este cenário torna-se cada vez mais viável.

Diante deste contexto complexo da utilização de mídias audiovisuais, surge a discussão teórica-metodológica do Design Audiovisual (DA). Inicialmente proposta na pesquisa de Becker, Gambaro e Ramos (2017) o DA integra aspectos de Mídia e Recepção da Comunicação a ferramentas e conceitos da Interação Humano Computador (IHC), com o intuito de auxiliar na descrição, análise e criação de sistemas audiovisuais. Em uma segunda publicação o DA é caracterizado pelos autores como uma arquitetura de conteúdo e interação capaz de atender a contextos de historicidade e generatividade, isto é, cenários em que tanto a interação quanto os dados do usuário são utilizados pelo sistema na oferta de conteúdo (TOSCANO et al., 2017).

2 OBJETIVO

Com o intuito de avançar a proposição dos estudos citados, este trabalho propõem a aplicação experimental do Design Audiovisual como a arquitetura de um projeto generativo, ou seja, um sistema que integra dados do usuário e interação para a geração de conteúdo. Tal estudo está sendo realizado pelos pesquisadores do Grupo de Pesquisa Design Audiovisual (GPDA) junto ao Laboratório de Aplicações em Vídeo Digital (LAVID) no âmbito das atividades de um projeto destinado ao Tribunal de Contas do Estado da Paraíba.

3 BASES TEÓRICAS E CONCEITUAIS

O desenvolvimento da capacidade de processamento computacional, bem como o surgimento de novos sensores permitiram aos softwares formas inovadoras de capturar e interpretar dados, isto é, compreender melhor o que significa a informação que está sendo capturada, processada e reproduzida (HANJALIC; LI-QUN XU, 2005; HAVALADAR; MEDIONI, 2010; HUGHES et al., 2014; KYBARTAS; BIDARRA, 2017).

Em paralelo aos estudos da computação é possível identificar uma crescente discussão acerca de diferentes modos de interação e fruição, presente na comunidade de pesquisadores dos estudos de mídia. Um cenário de múltiplos meios e formas de recepção (JENKINS; FORD; GREEN, 2014); em que os sistemas também atuam como parceiros do discurso (MANOVICH, 2013); e revelam o surgimento de novos padrões e convenções no processo de comunicação, representação e fruição audiovisual (MURRAY, 2011).

Assim como os estudos computacionais aproximam-se do valor semântico de seus dados, produtores e artistas do meio audiovisual também vem incorporado na criação de conteúdo o uso de dados e contextos diversos de interação para aprimorar suas premissas narrativas e comunicacionais.

3.1 Customizar, adaptar e personalizar

Os termos customização, adaptação e personalização são retratados na literatura de modo ligeiramente distintos. A lógica da customização aparece em contextos de oferta de recursos aos indivíduos para que estes possam ajustar elementos pré-estabelecidos conforme seus interesses. Já o termo adaptação é retratado em sua maioria a contextos técnicos como adequação a requisitos de rede, dimensão de tela ou acessibilidade. Por fim, a ideia de personalização está normalmente associada a filtragem de informações sobre a audiência de tal modo a recomendar, reposicionar, inserir, excluir, exibir determinado conteúdo ou modo de interação (AMY SCHADE, 2014; BARBOSA; SILVA, 2010; ELMAGARMID, 1997; GUERRINI et al., 2017; MEIXNER, 2017; XU; JIN; LUO, 2008).

Tendo compreendido a ideia de personalização, se faz necessário entender o que pode ser personalizado. Kybartas e Bidarra (2017) conceituam que a geração de elementos para personalização da experiência narrativa audiovisual passam por níveis estruturais distintos. O elemento mais geral é a narrativa, que é exatemente a forma ou expressão encontrada para apresentar uma história. Já a história é conjunto de elementos que constituiem uma relação linear. As histórias são compostas por tramas que por sua vez são compostas por eventos dramáticos. Uma história pode ser contada a partir de diversas formas narrativas, do mesmo modo que uma mesma história pode ter tramas variáveis que permitam a audiência multitramas e percepções. Essa abordagem estruturalista apresentada por Kybartas e Bidarra (2017) nos ajuda a compreender os níveis em que a ação dos sistemas digitais podem atuar de modo a impactar o conteúdo e a fruição dos indivíduos.

4 PROCEDIMENTOS METODOLÓGICOS

A arquitetura do Design Audiovisual é a base conceitual e metodológica deste estudo. O DA descreve essencialmente quatro funções, ou papéis, que os indivíduos podem assumir enquanto consomem conteúdo audiovisual: Audiência, que representa uma fruição passiva dos conteúdos; Sintetizador, onde os indivíduos organizam, compartilham e comentam conteúdos em redes sociais; Modificador, que corresponde às pessoas que se apropriam de conteúdos e os modificam; e Produtor, que representa os indivíduos, empresas e software que geram conteúdos novos. Cada papel tem seus níveis elevados, descritos como *Player*, quando estes se envolvem e usam todos os recursos disponíveis, ou, ainda, alteram totalmente os propósitos do conteúdo e do sistema.

Para o Design Audiovisual as relações de interação inerentes a esses níveis ocorrem por meio de quatro Linhas de Design: (a) Conteúdo; (b) Identidade, que baseia a relação pessoal com o conteúdo; (c) Motivação, que leva ao engajamento; e (d) Experiência, que relaciona a fruição à utilização avançada de recursos, que devem proporcionar uma experiência prática ou simbólica com o conteúdo e interfaces (ver apêndice A).

As Linhas de Design tem o intuito de auxiliar o desenvolvedor do sistema audiovisual (conteúdo e software) na elaboração do projeto de interação, engajamento e conteúdo de sua produção a partir de variáveis que impactam a fruição dos indivíduos.

Outro conceito relevante para a arquitetura do DA é a criação de gatilhos, que surgem das Linhas de Design do projeto. Um gatilho é qualquer elemento adicionado ao longo da interação com o intuito de estimular ações dos indivíduos. Podem vir como elementos sonoros, visuais, de interface, interação ou mesmo chamadas de ação no próprio conteúdo.

O DA será utilizado como base de descrição e criação deste trabalho à medida que propõe um conjunto de variáveis ou camadas relacionais que devem ser consideradas por produtores de conteúdo dentro de um contexto de uso de mídia complexo. De modo geral, as etapas que compõem o desenvolvimento a partir da arquitetura do Design Audiovisual, são:

- Identificar objetivo, principais usos do sistema e personas;
- Delimitar Linhas de Design;
- Projetar Gatilhos e Affordances;
- Desenvolver sistema (conteúdo e interface);[1]
- Testar o sistema.

Nas seções seguintes detalha-se a aplicação destes conceitos na construção do sistema audiovisual generativo.

5 RESULTADOS

O sistema apresentado neste trabalho compõe um projeto que integra estudantes da rede pública em um conjunto de atividades de produção, visita e acompanhamento das ações do TCE-PB.

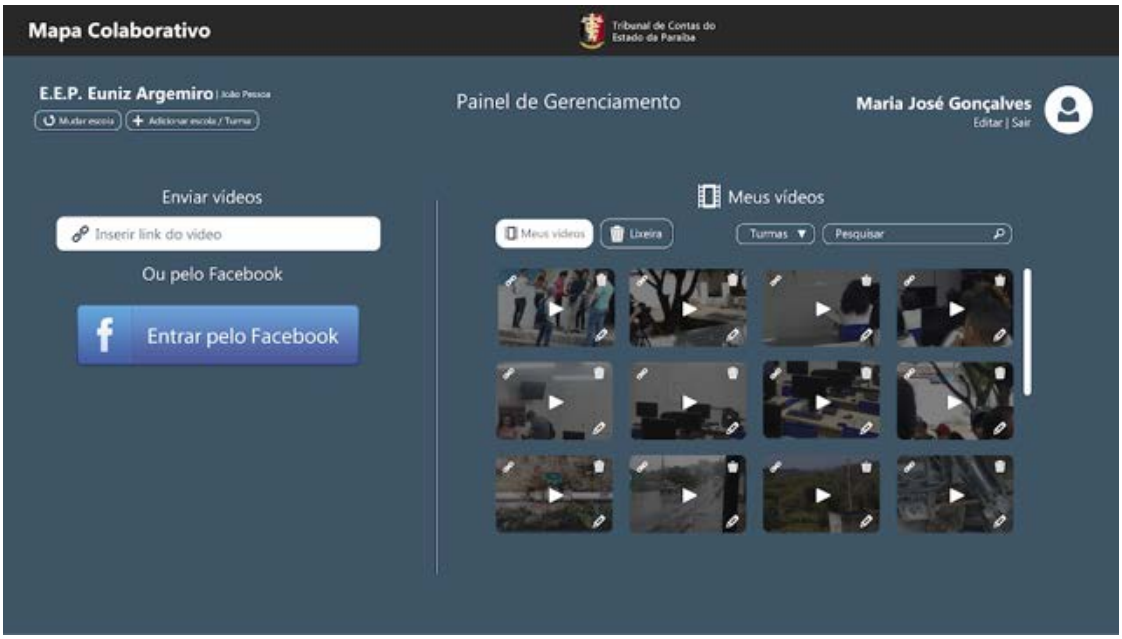
1 Nesta etapa são utilizados processos e ferramentas comuns a produção de vídeo e engenharia de software.

Neste cenário alunos da rede pública são convidados pelos professores a produzir vídeos sobre casos de sucesso ou aspectos ineficientes de sua comunidade escolar.

O conteúdo audiovisual que resulta desta ação tem como destino uma página do TCE-PB na mídia social Facebook a partir do perfil de cada estudante. No ato da publicação do material, cada aluno descreve o nome de sua escola, cidade e uso de uma *hashtag* pré-determinada. Após essa etapa, realizada pelos alunos, o professor responsável pela turma vai até um portal do TCE-PB, e com o seu login do Facebook tem a opção de revisar os conteúdos que foram identificados pelo sistema via a API GRAPH[2] conforme apresentado na figura 1.

Essa coleta de dados (vídeo, nome do usuário e escola) serve a dois sistemas, são eles: sistema de visualização avançada e colaborativa via mapa da Paraíba; sistema de apoio a palestras, também denominado de gerador de vídeos personalizados.

Figura 1. Painel de Gerenciamento



Dados da Pesquisa, 2018

5.1 Sistema de vídeos personalizados

O TCE-PB recebe mensalmente estudantes da rede pública para conhecer sua estrutura física e atuação na sociedade. Como uma das estratégias de integrar esses visitantes na temática do controle social e educação fiscal, propõe-se então um sistema de apoio a palestras que tem o objetivo de integrar produtos audiovisuais institucionais a produção colaborativa dos estudantes, com o intuito de personalizar a experiência de cada turma que visita o TCE-PB.

2 API disponibilizada pelo Facebook para a utilização de informações do usuário em serviços web externos.

Aplicando a arquitetura do Design Audiovisual na construção deste sistema temos a seguinte proposta: a partir de um vídeo institucional, que define a linha de Conteúdo, e que conceitua a importância da população nos atos de fiscalização e acompanhamento da gestão pública, os vídeos dos alunos são inseridos como elementos integrante a narrativa (Linha de Identidade e Motivação), conforme a figura 2.

O vídeo personalizado, isto é, gerado com base nos dados dos estudantes dispõe da seguinte estrutura:

Parte 1: Contextualização sobre aspectos da má gestão e corrupção, e como a participação popular é importante para combater e acompanhar as falhas na gestão pública. Seguindo a arquitetura do DA, nesta etapa do vídeo devem ser trabalhados recursos narrativos (ritmo de edição, trilha sonora, formas de expressão, entre outros) que sejam coerentes com os estudantes de tal modo a estimular a atenção e interesse pelo tema.

Parte 2: Exemplos de atitudes de controle social a partir dos vídeos de alunos. Seguindo a arquitetura do DA, a utilização dos vídeos dos próprios alunos atua como um Gatilho da Inércia, um estímulo à atenção e formação de Identidade e afinidade com o conteúdo.

Figura 2. Abstração da personalização narrativa

Dados da Pesquisa, 2018

Parte 3: Especialistas explicam que os atos dos estudantes representam o controle social, sendo este fundamental para a sociedade. Seguindo a arquitetura do DA, nesta etapa do vídeo utiliza-se os princípios da Linha de Conteúdo e Identidade, uma vez que a participação dos alunos é tratada a nível de conteúdo como parceira do TCE-PB.

Parte 4: Crédito do filme com o nome dos participantes (especialistas, escola e alunos). Esta etapa é um Gatilho da Inércia, uma vez que atua como estímulo a atenção e formação de Identidade e afinidade com o conteúdo.

Parte 5: Disponibilização de QR code para compartilhamento do vídeo. Seguindo a arquitetura do DA, esta etapa atua como um Gatilho de Ação, que a partir da Identidade estimula o estudante (Audiência) a alternar o papel para Sintetizador, ao compartilhar o material em suas próprias mídias sociais.

Outras linhas de design que podem ser impactadas pelo sistema são a Motivação e Experiência. Uma vez que este material personalizado a partir dos dados dos indivíduos é disponibilizado em um formato de arquivo compatível com mídias sociais (Youtube, Facebook ou Whatsapp) é possível viabilizar cenários em que os estudantes possam modificar, editar ou mesmo remixar o conteúdo.

Para garantir tal cenário de fruição e interação, foram delimitados os seguintes requisitos funcionais para o sistema: Estabelecer conexão com o Facebook via API Graph; Incorporar *hashtag*, vídeo, nome do estudante e escola a partir de uma postagem pública no Facebook; Inserir os dados importados dentro do projeto vídeo "Controle Social"; Renderizar todo o material (vídeos institucionais fixos, personalizáveis e dados) de modo a gerar um único arquivo de vídeo que seja compatível com HTML5 e serviços como youtube e vimeo;

6 CONSIDERAÇÕES FINAIS

Este trabalho se propôs experimentar a arquitetura do DA como base para a construção de um sistema generativo. A possibilidade de adaptar o conteúdo instaura um cenário emergente para os estudos de mídia, a criação de conteúdo para fruição individual ou social (pequenos grupos de pessoas) ao invés do tradicional modelo Broadcasting.

A pesquisa sobre a construção de sistemas audiovisuais generativos ainda se encontra em andamento, tendo como próxima etapa a implementação dos requisitos, realização de testes com usuário para validação da usabilidade e engajamento. Todavia, até o presente momento foram aprendidas as seguintes lições:

- As quatro linhas de Design e gatilhos do Design Audiovisual foram úteis no auxílio do levantamento de requisitos funcionais do conteúdo e interfaces interação;
- A personalização de trechos do vídeo apresenta-se como elemento potencial para projetar inércia e ação perante os níveis de engajamento e dispositivos ofertados.

Os impactos destes elementos ficam a cargo de estudos futuros, na continuação desta pesquisa.

REFERÊNCIAS

AMY SCHADE. **Video Usability**. Disponível em: <https://www.nngroup.com/articles/video-usability/>. Acesso em: 7 jul. 2018.

BARBOSA, S. D. J.; SILVA, B. S. DA. **Interação Humano-Computador**. 1. ed. Rio de Janeiro: Elsevier, 2010.

BECKER, V.; GAMBARO, D.; RAMOS, T. S. Audiovisual design and the convergence between HCI and audience studies. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [s.l: s.n.]. v. 10271p. 3–22.

ELMAGARMID, A. K. Smart Video Text : An Intelligent Video Database System. 1997.

GUERRINI, F. et al. Interactive Film Recombination. **ACM Transactions on Multimedia Computing, Communications, and Applications**, v. 13, n. 4, p. 1–22, 12 ago. 2017.

HANJALIC, A.; LI-QUN XU. Affective video content representation and modeling. **IEEE Transactions on Multimedia**, v. 7, n. 1, p. 143–154, fev. 2005.

HAVALADAR, P.; MEDIONI, G. **Multimedia Systems: Algorithms, Standards, and Industry Pratices**. 1. ed. Boston,MA: Course Technology: CENGAGE Learnig, 2010.

HUGHES, J. F. et al. **Computer Graphics: Principles and Practice**. Third Edit ed. Ohio: Addison- Wesley, 2014.

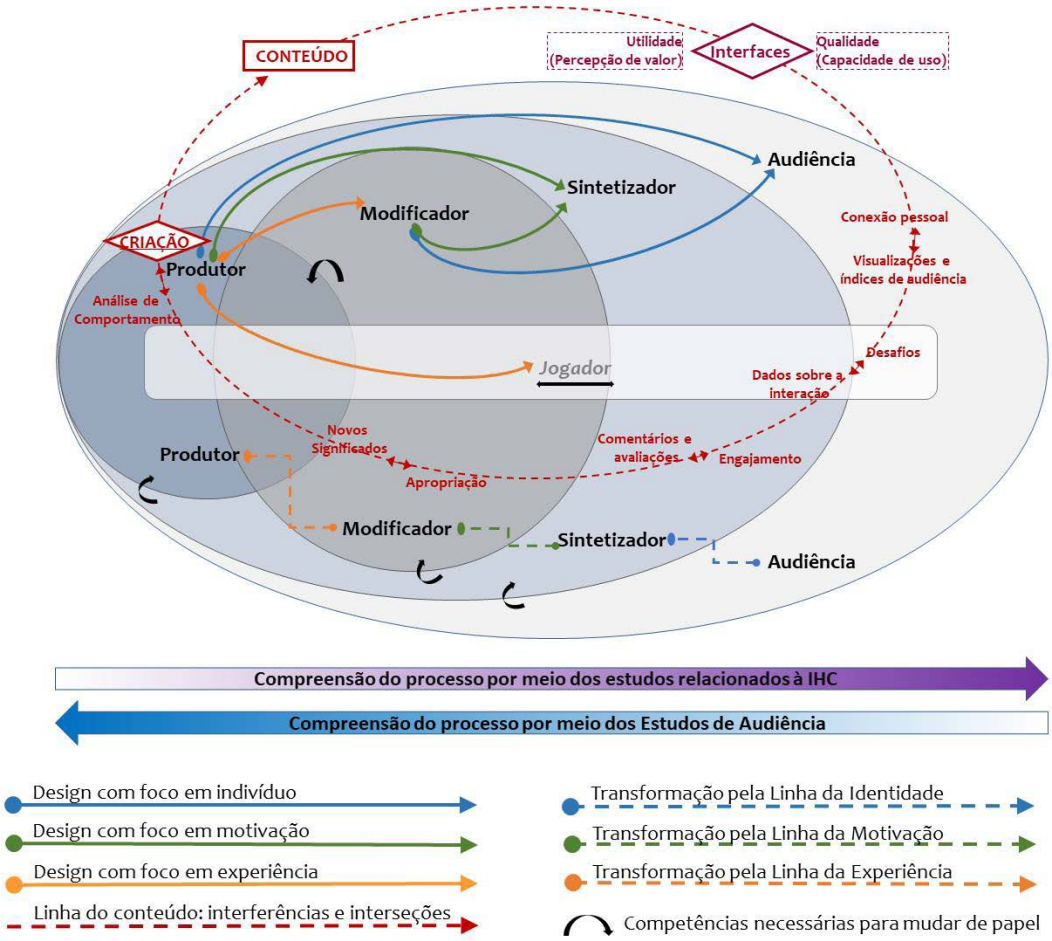
KYBARTAS, B.; BIDARRA, R. A Survey on Story Generation Techniques for Authoring Computational Narratives. **IEEE Transactions on Computational Intelligence and AI in Games**, v. 9, n. 3, p. 239–253, 2017.

MEIXNER, B. Hypervideos and Interactive Multimedia Presentations. **ACM Computing Surveys**, v. 50, n. 1, p. 1–34, 2017.

TOSCANO, R. et al. Arquitetura de design colaborativo para imersão temporal e espacial em vídeos de altíssimas resoluções e HFR. In: **O futuro da videocolaboração: perspectivas**. Porto Alegre, RS: Sociedade Brasileira de Computação (SBC), 2017.

XU, M.; JIN, J. S.; LUO, S. Personalized video adaptation based on video content analysis. **Proceedings of the 9th International Workshop on Multimedia Data Mining held in conjunction with the ACM SIGKDD 2008 - MDM '08**, p. 26–35, 2008.

APÊNDICE A – ARQUITETURA DO DESIGN AUDIOVISUAL



Fonte: (TOSCANO et al., 2017)

DIREITO AUTORAL + CREATIVE COMMONS: uma solução para a academia na Sociedade da Informação?

COPYRIGHT + CREATIVE COMMONS:
a solution for academics in the Information Society Era?

Enrique Muriel-Torrado[1], Rebeca Silva Fernandes de Moura [(N)]
(1) Universidade Federal de Santa Catarina (UFSC), Programa de Pós-Graduação em Ciência da Informação (PGCIN), R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900, enrique.muriel@ufsc.br.
(N) Universidade Federal de Santa Catarina (UFSC), Programa de Pós-Graduação em Ciência da Informação (PGCIN), R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900, rebeca.andrade@ufsc.br.

Resumo:
O Direito Autoral visa proteger os autores e suas criações e buscar a disseminação do conhecimento na sociedade da informação, porém, é muitas vezes interpretado como uma lei restritiva. Desta forma, surgem alternativas de flexibilização, padronização e utilização das obras por terceiros, as licenças *Creative Commons* que permitem que o autor escolha a maneira que deseja que sua obra seja propagada, tornando possível que a sociedade compartilhe e consulte obras disponíveis no ambiente digital. As licenças objetivam incentivar o compartilhamento de obras de maneira organizada, seguindo o pré-estabelecido pela lei de nº 9.610, que consolida e regulariza os direitos autorais no Brasil. Neste artigo, discutiremos se as licenças *Creative Commons* podem ser utilizadas para substituir uma lei que encontra-se desatualizada, baseando em algumas atividades comumente realizadas pelos docentes universitários na Sociedade da Informação. A metodologia utilizada foi através de levantamento bibliográfico acerca da matéria e da proposição de lista de ações que ocorrem no ambiente educacional e quais direitos atuam sobre elas. Como conclusão, percebe-se que a lei de Direitos Autorais atual não engloba todas as necessidades do usuário, sendo as licenças capazes de atuar como alternativa de proteção aos autores e oferecer aos usuários alguns direitos, mas que não pode nem deve substituir a lei.
Palavras-chave: Direito Autoral. Creative Commons. Sociedade da Informação.

Abstract:
The copyright aims to protect the authors and their creations, as well as communicate knowledge in the Information Society. Despite of its goals, it is perceived as hard and prohibited. Thus, flexible alternatives rise, along with standardization and use by third parts, the named Creative Commons licenses, allowing the authors to chose the rules that govern how the work is used and publicized, making possible that general society shares and makes use of available digital work. The goal for this licenses is subsidize the organized usage, under the Law 9610 guidelines, which stablishes and regulates Copyright Ownership in Brazil. This work discusses if Creative Commons licenses can be used in substitution to an old law, based in some common activities used by the faculty members within the Information Society. The methods involved a systematic review and a proposal for actions that occur in the educational environment and which rights are involved. In conclusion, the current Authorship Law does not cover the needs from the users, and the Creative Commons licenses are a support and an alternative for author's protection and can bring a few rights, but these licenses should not be used to substitute the actual Law.
Keywords: Copyright. Creative Commons. Information Society.

1 INTRODUÇÃO

O direito autoral sofre constantes mudanças ocasionadas pela transformação da própria sociedade, sociedade essa definida por Werthein (2000), como Sociedade da Informação (SI), que destaca que a expressão sociedade da informação passou a ser utilizada nos últimos anos deste século. Devido as transformações ocorridas, faz-se necessário que o direito autoral se adapte para que seja possível proteger as obras de maneira eficaz.

Coutinho e Lisbôa (2011) relacionam o conceito de SI ao de uma sociedade que vive em constante mutação ocasionada pelo avanço na ciência e na tecnologia.

A tecnologia é capaz de proporcionar maior possibilidade de acesso às informações e torná-las disponíveis a terceiros de maneira mais rápida e eficiente. Ao oferecer diversas possibilidades, seja no acesso ou no compartilhamento de informações, a tecnologia se torna cada vez parte integrante do dia a dia dos membros da sociedade da informação.

Pelo fato da tecnologia se fazer tão presente em nossas atividades diárias é possível considerar que o surgimento da internet reorganizou a forma de viver em sociedade e alterou a forma de comunicação drasticamente, transformando bruscamente a sociedade (BURCH, 2005).

O aumento exponencial de possibilidades de acesso à internet e de oferta de tecnologia contribuem com a divergência entre às necessidades demandadas pela população há 20 anos, quando a Lei brasileira de Direitos Autorais (LDA) foi promulgada, com as necessidades demandadas atualmente pela sociedade da informação, pois a tecnologia contribui para que o conhecimento flua e torne-se ultrapassado em velocidade muito maior que anteriormente.

Na atualidade algumas correntes apontam que existe maior necessidade de proteção das obras ao considerarmos que elas são disponibilizadas e acessadas facilmente no ambiente digital. Desta forma, faz-se necessário que tanto as obras de autores que desejam compartilhá-las, como aquelas cujo autor não deseja tal difusão no ambiente digital, sejam resguardadas legalmente de modo a proteger os autores de utilização indevida de suas criações por terceiros.

O direito autoral é definido por Duarte e Pereira (2009) como:

> "[...] o poder que o autor, o criador, o tradutor, o pesquisador ou o artista tem de controlar o uso que se faz de sua obra. Basicamente os direitos autorais trabalham com a imaterialidade, sendo esta, a principal particularidade da propriedade intelectual."

No Brasil o direito autoral está consolidado à lei de nº 9.610, aprovada em 19 de fevereiro de 1998 (BRASIL, 1998), que surgiu para, conforme seu próprio preâmbulo, normatizar, consolidar e atualizar leis brasileiras já existentes. A lei específica que regula a matéria é fruto da evolução de diversas leis brasileiras e tratados internacionais, como apresentado no apêndice A.

Conforme Dias e Oliveira (2015) o direito autoral nasce exatamente no momento de criação da obra, ou seja, quando uma obra é cadastrada em suporte material em concomitância nasce também o direito autoral para proteger e garantir que todos os direitos sejam reservados ao autor.

O direito autoral busca proteger os autores, e frequentemente é interpretado como elemento restritivo. Na tentativa de solucionar essa perspectiva restritiva e buscar alternativas de flexibilização de uso das obras surgem as licenças *Creative Commons*.

As licenças *Creative Commons* configuram-se como uma camada de proteção sobre a LDA ao permitir a utilização de obras, protegidas por direito autoral, de maneira flexível oferecendo ao autor o poder de escolher como compartilhar suas criações ao mesmo passo que explica aos usuários como utilizá-las respeitando as condições de uso impostas pela licença escolhida pelo autor.

Segundo Muriel-Torrado (2012) as licenças *Creative Commons* são as licenças mais conhecidas e desenvolvidas a nível internacional, além de buscar, como ideia principal, contribuir com o acesso e compartilhamento da informação ao público.

Importante destacar que as licenças *Creative Commons* indicam tipos diferentes de permissões de utilização das obras, ficando a critério do autor restringir mais ou menos a utilização da mesma por terceiros.

2 OBJETIVOS

No presente artigo caracterizaremos a sociedade da informação, bem como o direito autoral e buscaremos alternativas flexíveis de uso das obras, através das licenças *Creative Commons*.

Como objetivo geral da pesquisa pretende-se identificar se algumas das principais atividades desenvolvidas pelos acadêmicos, que entram em conflito com os direitos autorais, poderiam ser sanadas com a utilização das licenças *Creative Commons*.

Busca-se também identificar se as licenças *Creative Commons* são suficientes para que os professores possam exercer essas atividades, substituindo assim a LDA ou se é necessário continuar modificando a lei para atender as demandas dos profissionais da docência.

3 PROCEDIMENTOS METODOLÓGICOS

O procedimento metodológico é uma pesquisa bibliográfica, exploratória, qualitativa e descritiva. Começando pela elaboração de uma breve listagem atividades exercidas no ambiente educacional *on-line* pelos professores universitários (baseado em MURIEL-TORRADO, 2012), para posteriormente relacionar essas atividades exercidas nas atividades docentes com os direitos autorais que estão por trás e assim identificar se poderíamos considerar as licenças *Creative Commons* como substitutas da LDA, claramente desatualizada na sua aplicação no contexto da Sociedade da Informação.

4 RESULTADOS

Entre as atividades docentes que desempenha um professor nas suas funções, podemos destacar (baseado em MURIEL-TORRADO, 2012): 1) Fazer upload de obras intelectuais no *Moodle* ou plataformas similares; 2) envio de artigos ou outras obras protegidas por e-mail para os alunos da turma ou companheiros; 3) Baixar obras da internet; 4) Fotocopiar ou digitalizar obras; 5) ou compartilhar fotocópias de obras com os alunos.

Segundo a Lei Brasileira de Direitos Autorais, para a primeira ação os direitos afetados seriam: reprodução e comunicação pública, já que primeiro subimos uma cópia da obra e depois a disponibilizamos para um conjunto de pessoas sem cópias físicas; no caso do envio de artigos por e-mail para uma turma, além da reprodução da obra efetuada para subir no e-mail, igual do que no anterior, se produz uma comunicação pública da mesma; para realizar um *download* de arquivos da internet afeta-se o direito de reprodução das obras; no caso de fotocopiar ou digitalizar arquivos tem-se o direito de reprodução afetado; e, por fim, para compartilhar fotocópias de obras com a turma afeta-se o direito de distribuição, pois seria em um suporte físico tangível.

Para cada ação desenvolvida e seu direito afetado, apresentados acima, nota-se que só é possível realizá-la dentro das exceções e limitações apresentadas no art. 46 da LDA ou com a permissão do titular dos direitos. Estas possibilidades restringiriam significativamente a utilização, por exemplo, de obras completas, embora fosse para fins de estudo. Mas se essas mesmas obras estivem licenciadas com *Creative Commons* (com quaisquer das suas opções) não feriria os legítimos direitos dos titulares dos direitos. Desta forma, as obras poderiam ser compartilhadas, subidas no *Moodle*, baixadas, enviadas por qualquer meio digital ou fotocopiadas, isto é, utilizadas de maneira muito mais flexível, de acordo com a licença adotada pelo autor e sem precisar solicitar nenhuma permissão.

Cabe ressaltar que a lei brasileira de direitos autorais em seu art. 68 (BRASIL, 1998) esclarece que sem autorização prévia e expressa do autor suas criações não poderão ser utilizadas. Desta forma, a comunicação pública, seja ela no ambiente educacional ou não, depende de prévia autorização do autor ou titular da obra, caso contrário, há infração dos direitos de autor.

4 CONCLUSÃO OU CONSIDERAÇÕES FINAIS

Após análise é possível notar que a lei de direitos autorais brasileira, por si só, não contempla as necessidades demandadas pela atual sociedade da informação, particularmente no que se refere à reprodução, tradução e compartilhamento da informação científica no ambiente digital, nem mesmo mediante as exceções e limitações recolhidas no Capítulo IV da lei.

Por este motivo surgiram alternativas complementares que tornam possível a proteção dos autores e ao mesmo tempo que oferece aos usuários alguns direitos e flexibilidade na utilização das obras. Isto se articula principalmente mediante as licenças *Creative Commons*, que permitem que os autores possam compartilhar os direitos que considerem oportunos e se reservar outros, conforme sua própria escolha.

Embora as licenças *Creative Commons* sejam uma excelente ferramenta, devemos lembrar que a LDA protege às obras desde o ato de criação, não havendo exigência de que as obras sejam registradas para que sejam protegidas. Por outro lado, utilizar as licenças *Creative Commons* é um ato voluntário, aonde o autor precisa licenciar sua obra e decidir qual tipo de licença irá atribui-la. Embora as licenças sejam muito importantes, precisam obrigatoriamente de uma LDA para poder existir, e essa mesma lei precisa também de reformas adequadas para melhorar seus fins. Mas, as licenças oferecem uma flexibilidade fundamental para otimizar o trabalho de docentes e alunos, requerendo apenas um pequeno aprendizado para desfrutar das mesmas. Por outro lado, seria interessante uma estratégia conjunta de melhora da legislação para esses fins e a implantação de uma política pública no

Ministério de Educação orientada à formação assim como uma aplicação padrão de licenças *Creative Commons* nos materiais utilizados no ensino no Brasil como um todo, para que professores e alunos possam utilizar materiais educativos no ambiente digital dentro da legislação.

REFERÊNCIAS

BRASIL. **CONSTITUIÇÃO DA REPÚBLICA DOS ESTADOS UNIDOS DO BRASIL, DE 24 DE FEVEREIRO DE 1891.** Organiza um regime livre e democrático, 1891. Disponível em: < http://www.planalto.gov.br/ccivil_03/constituicao/constituicao91.htm>. Acesso em: 20 set. 2018.

BRASIL. **CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL, DE 05 DE OUTUBRO DE 1988.** Institui um Estado Democrático, 1988. Disponível em: < http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm>. Acesso em: 10 set. 2018.

BRASIL. **DECRETO Nº 4.790, DE 2 DE JANEIRO DE 1924.** Regula os direitos autorais e dá outras providências., 1924. Disponível em: <http://www2.camara.leg.br/legin/fed/decret/1920-1929/decreto-4790-2-janeiro-1924-565512-republicacao-89686-pl.html>. Acesso em: 12 set. 2018.

BRASIL. **DECRETO Nº 75.699, DE 6 DE MAIO DE 1975.** Promulga a Convenção de Berna para a Proteção das Obras Literárias e Artísticas, de 9 de setembro de 1886, revista em Paris, a 24 de julho de 1971. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/1970-1979/D75699.htm>. Acesso em: 12 set. 2018.

BRASIL. **LEI Nº 496, DE 01 DE AGOSTO DE 1898.** Define e garante os Direitos Autorais, 1898. Disponível em: <http://www2.camara.leg.br/legin/fed/lei/1824-1899/lei-496-1-agosto--1898-540039-publicacaooriginal-39820-pl.html>. Acesso em: 20 set. 2018.

BRASIL. **LEI Nº 3.071, DE 01 DE JANEIRO DE 1916.** Código Civil dos Estados Unidos do Brasil, 1916. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L3071.htm>. Acesso em: 20 set. 2018.

BRASIL. **LEI Nº 5.988, DE 14 DE DEZEMBRO DE 1973.** Regula os direitos autorais e dá outras providências., 1973. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L5988.htm>. Acesso em: 21 set. 2018.

BRASIL. **LEI Nº 9.609, DE 19 DE FEVEREIRO DE 1998.** Dispõe sobre a proteção da propriedade intelectual de programa de computador, sua comercialização no País, e dá outras providências., 1998. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9609.htm>. Acesso em: 10 set. 2018.

BRASIL. **LEI Nº 9.610, DE 19 DE FEVEREIRO DE 1998.** Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências., 1998. Disponível em: <http://www.planalto.gov.br/ccivil_03/LEIS/L9610.htm>. Acesso em: 10 set. 2018.

BURCH, Sally. **Sociedade da informação versus Sociedade do conhecimento.** In: Desafios de Palavras: Enfoques Multiculturais sobre as Sociedades da Informação, nov., 2005. Disponível em: < https://vecam.org/archives/article519.html>. Acesso em: 19 set. 2018.

COUTINHO, Clara; LISBÔA, Eliana. **Sociedade da Informação, do Conhecimento e da Aprendizagem: desafios para educação no século XXI.** 2011. Disponível em: <http://revista.educ.ie.ulisboa.pt/arquivo/vol_XVIII_1/artigo1.pdf>. Acesso em: 17 set. 2018.

DIAS, Guilherme Ataíde. OLIVEIRA, Bernardina M. J. Freire de. **Informação, Direito Autoral e Plágio.** João Pessoa, 2015.

DUARTE, Eliane Cordeiro de Vasconcellos Garcia; PEREIRA, Edmeire Cristina. **Direito Autoral: Perguntas e Respostas.** Curitiba: UFPR, 2009.

DUDZIAK, Elisabeth Adriana. **Licenças Creative Commons: saiba mais sobre isso.** 2016. Disponível em: <http://www.sibi.usp.br/?p=5763> Acesso em: 20 set. 2018.

MURIEL-TORRADO, Enrique. **Los derechos de autor y la enseñanza en la universidad: el papel de la biblioteca universitaria.** 2012. Tese (Doctorado en Documentación) – Facultad de Comunicación y Documentación, Universidad de Granada, Espanha.

WERTHEIN, Jorge. **A sociedade da informação e seus desafios.** Brasília, mai/ago. 2000. Disponível em: <http://www.scielo.br/pdf/ci/v29n2/a09v29n2.pdf>. Acesso em: 19 set. 2018.

Apêndice A – Evolução das Leis de Direito Autoral no Brasil.

Lei	**Abrangência**
Constituição Federal de 1891	Foi a primeira legislação a tratar do tema de direito autoral e, mesmo tendo sido promulgada dois anos após o nascimento da República, já se preocupava com os direitos relacionados aos autores.
Lei nº 496 de 1898	Foi a primeira lei específica a regular a matéria e ela define e garante os direitos autorais no país. Posteriormente complementada com a Convenção de Berna que aconteceu em 1886.
Lei nº 3.071 de 1916 (Código Civil)	Regulamentou o direito autoral nos arts. 649 a 673. Fez com que a lei nº 496 perdesse sua eficácia e fosse revogada por completo.
Decreto nº 4.790 de 1924	Esse decreto instituiu o direito autoral no Brasil, preocupando-se com composições teatrais e musicais.
Lei nº 5.988 de 1973	Regula direitos de autor e direitos que lhe são conexos e parte dessa legislação vigorou até 1998 com lei atual de direito autoral.
Decreto nº 75.699 de 1975	Ratifica a convenção de Berna (Suíça).
Constituição Federal de 1988	A CF expõe em seu art. 5º que aos autores pertence o direito exclusivo de utilização, publicação ou reprodução de suas obras, transmissível aos herdeiros pelo tempo que a lei fixar.
Lei nº 9.609 de 1998	Essa lei dispõe sobre a proteção da propriedade intelectual de programa de computador, sua comercialização no País. Sendo conferido às obras literárias pela legislação de direitos autorais e conexos vigentes no País.
Lei nº 9.610 de 1998	Essa lei, que se encontra atualmente vigente, veio para ampliar os direitos e garantias dos autores de obras.

Fonte: Elaborado pelos autores – 2018.

DIREITO DE PRIVACIDADE NO BRASIL:
Tendências

PERSONAL DATA PRIVACY IN BRAZIL:
Tendencies

Rossana Gleucy de Ávila Chagas e Carvalho[1], Márcio Matias[2]
(1) Universidade Federal de Santa Catarina, R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900 - professoraebd@gmail.com
(2) Universidade Federal de Santa Catarina, R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900 – Matias97@gmail.com

Resumo:
A proposta deste artigo é constatar as tendências que uma decisão do Supremo Tribunal Federal pode promover e a nova Lei Geral de Proteção de Dados Pessoais, em relação à privacidade de dados pessoais na doutrina e jurisprudência brasileiras. Para tanto foi feita uma pesquisa qualitativa nos argumentos doutrinários e jurisprudenciais utilizados pela Ministra Cármem Lúcia, relatora, na análise do mérito da Ação Direta de Inconstitucionalidade nº 4.815/DF, ajuizada em 5 de julho de 2012, pela Associação Nacional dos Editores de Livros – ANEL, e considerações sobre a efetividade da Lei Geral de Proteção de Dados Pessoais. Os resultados encontrados na pesquisa apontam para a desvalorização do direito à privacidade, intimidade e identidade em relação a outros princípios e direitos tais como: a liberdade de expressão, pensamento e opinião; direito de informação; princípio da transparência na gestão pública dentre outros. A conclusão a que se chegou é a de que o panoptismo idealizado por Bentham e destacado por Foucault não está, nos dias atuais, ligado apenas a atividades estatais. Há ainda interesses corporativos ou comerciais e pessoais que vêm enfraquecendo o direito constitucional à privacidade.
Palavras-chave: Privacidade. Liberdade de expressão. Panoptismo.

Abstract:
The objective of this article is to identify trends that a decision of the Federal Supreme Court and the General Law for Personal Data Protection could cause to the privacy into Brazilian legal doctrine and jurisprudence. For this purpose, this qualitative research took place about doctrine and jurisprudence used by the Minister Reporter Cármem Lúcia at the examination of the merits by the Direct Action of Unconstitutionality nº 4.815/DF, filed on July 5th, 2012, by the pela Associação Nacional dos Editores de Livros – ANEL and about the effectiveness of the General Law for Personal Data Protection. The results found in this research indicate devaluation of the right privacy against other fundamental rights in Brazil, such as freedom of expression, thought, opinion, right to information, transparency in the public management, among others. The conclusion reached in this research is that the panopticon idealized by Bentham and emphasized by Foucault is not, nowadays, connected to state-activities. There are private corporate, commercial and personal interests that are weakening the constitutional right to privacy.
Keywords: Privacy. Freedom of expression. Panopticism.

1 INTRODUÇÃO

O direito a privacidade sofreu transformações ao longo do tempo (CANCELIER, 2017, p. 214). Do "right to be alone" (direito a ser deixado só) conceito de Brandeis (1890) até o sentido mais pragmático da concepção atual "caracterizada pela liberdade de autodeterminação informativa, isto é, a capacidade de controlar as informações pessoais pelo seu titular" (MACHADO, 2014, p. 338).

No entanto, nos dias atuais, o controle do indivíduo sobre suas informações pessoais não se apresenta como uma tarefa fácil. Os avanços das ferramentas da tecnologia da informação e da comunicação ocorridos nas últimas décadas alteraram o comportamento as relações sociais de convivência humana.

Há dois fatores jurídicos que devem especialmente ser analisados quando se pretende analisar as tendências em relação ao direito à privacidade no Brasil: o primeiro, é a promulgação da Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709 de 13 de agosto de 2018), novidade no ordenamento jurídico brasileiro por tratar especificamente sobre o tema "privacidade" e que alterou o Marco Civil da Internet (Lei nº 12.965); Celso Lafer (2006, p. 241) afirma que a positivação de um direito humano por vezes exacerba os problemas práticos de sua tutela, em vez de eliminá-los, como normalmente se espera.

O segundo fator que precisa ser levado em consideração é a decisão da Suprema Corte brasileira, proferida na Ação Direta de Inconstitucionalidade – ADI nº 4815/DF, ajuizada em 5 de julho de 2012, pela Associação Nacional dos Editores de Livros – ANEL. Essa ação, levou o Supremo Tribunal Federal – STF a discutir o direito fundamental à privacidade e o processo foi relatado pela Ministra Cármem Lúcia.

Especificamente, esta ADI questionou a constitucionalidade da proibição contida nos artigos 20 e 21 do Código Civil, que exigem autorização para a publicação de biografias não autorizadas de notórias personalidades brasileiras, como se fosse uma espécie de censura privada.

A relevância jurídica desta decisão reflete-se no impacto social que dela se espera, pois, o voto da relatora foi aceito por unanimidade pelos outros ministros, além disso, como é sabido, após a decisão da máxima corte brasileira, "os tribunais devem uniformizar sua jurisprudência e mantê-la estável, íntegra e coerente", conforme o art. 926 do Código de Processo Civil de 2015.

2 OBJETIVOS

Assim, o objetivo do presente artigo é constatar as tendências em relação à privacidade, intimidade e identidade na doutrina jurídica e jurisprudência brasileiras diante da novidade na legislação (Lei de Proteção de dados pessoais) e do julgado do STF, respectivamente.

3 PROCEDIMENTOS METODOLÓGICOS

Para que o objetivo supracitado fosse alcançado foi feita uma pesquisa qualitativa organizada a partir dos argumentos doutrinários e jurisprudenciais utilizados pela Ministra Cármem Lúcia,

relatora da ADI nº 4815/DF e do texto da Lei Geral de Proteção de Dados no que tange aos direitos de privacidade, intimidade e identidade, em contraste com a ideia do panoptismo desenvolvido por Jeremy Bentham e destacado por Foucault.

Vale ressaltar que dezesseis entidades também se manifestaram no processo, tratando-se assim, de um colegiado de opiniões que foram consideradas pelo órgão, ouvidas as associações de escritores, órgãos públicos, universidade, etc.

No mesmo sentido, levou-se 8 anos para discutir e redigir alterações no projeto de lei que, posteriormente, gerou a Lei de Proteção de Dados Pessoais.

4 RESULTADOS

No que tange à análise da decisão na ADI 4815/DF, os resultados encontrados na pesquisa apontam para a desvalorização do direito à privacidade, intimidade e identidade em relação a outros princípios, direitos e deveres jurídicos tais como a liberdade de expressão e liberdade de conhecimento.

Castels (1999) chama a atenção para o fato de que "a geração, processamento e transmissão de informação torna-se a principal fonte de produtividade e poder" (Castells, 1999, p.21). Não obstante, percebe-se que, ao tratar do direito de liberdade de expressão, a Ministra Cármem Lúcia compreende essas relações de poder e suas consequências para a sociedade.

Assim, a tendência no STF, e consequentemente, na jurirsprudência brasileira, é a de reforçar o caráter comunitarista dos direitos da personalidade, no bojo dos quais encontram-se o direito à privacidade e à intimidade. Segundo essa linha de pensamento, dá-se mais importância ao interesse coletivo que ao interesse "egoísta" individual.

Em relação à Lei Geral de Proteção de Dados Pessoais, a opção do legislador pátrio é seguir as diretrizes da legislação correlata na Europa, destacando princípios importantes como o princípio da finalidade, adequação, necessidade e transparência, dos quais falaremos adiante, dentre outros. Faltando porém a criação da autoridade fiscalizadora da proteção de dados.

4.1 Direito de Liberdade de Expressão (art. 5º, IV e V).

O STF entende que o direito de liberdade de expressão deve ser entendido em seu sentido amplo, como sendo a fonte do direito de liberdade de todos os outros tipos de expressão do pensamento, abarcando a liberdade de opinião atividade intelectual, artística, científica, liberdade de conhecimento e de comunicação, mas também é uma liberdade de conteúdo espiritual, como a expressão musical e da comunicação pelas artes plásticas (STF, 2015).

Ou seja, o direito à privacidade tem intrínseca relação com a liberdade. Este também é o pensamento de Jabur (2000).

Em contrapartida, citando Canotilho, a Ministra ressalta que o direito de liberdade de expressão também envolve a liberdade de profissão, de livre iniciativa econômica, de prestação de serviços e de propriedade (STF, 2015). Defendeu ainda a expressão como sendo "responsável pela

construção e reconstrução de relações sociais, políticas e econômicas". É "o fio condutor de lutas e direitos fundamentais" (STF, 2015, p. 51).

O princípio da livre manifestação do pensamento, diz respeito à autonomia e ao livre desenvolvimento da personalidade do indivíduo (STF, 2015).

Numa dimensão social e política, princípio magno do Estado Democrático de Direito (STF, 2015). É condição e garantia da democracia, do pluralismo político e da tolerância. É, portanto, também, um direito político.

Cármem Lúcia, citando Wolfgang Sarlet, afirma que o direito de liberdade de expressão "assume uma espécie de posição preferencial", sem, contudo, lhe atribuir "a condição de direito absolutamente imune a qualquer limite e restrição sem se estabelecer uma hierarquia prévia entre as normas constitucionais" (STF, 2015, p.36). No entanto, a liberdade de expressão é reconhecida como que numa posição de vantagem no caso de conflito com outros bens fundamentais individuais e coletivos (STF, 2015).

Por entender as relações de poder envolvidas no exercício do direito de liberdade de expressão, a ministra entende que nem o Estado, nem particulares podem lhe impor limites, exceto quando o exercício desse direito atentar contra outros bens juridicamente protegidos ou para garantir a liberdade do outro (STF, 2015).

Assim, "conflitos entre particulares podem atingir direitos fundamentais pela desproporcionalidade, ou seja, por abuso ou exorbitância do poder exercido por um com relação a outro ou em contrariedade ao interesse público" (STF, 2015, p.58). Citando julgado da Corte Europeia de Direitos Humanos, a Ministra adverte que "toda formalidade, condição, restrição ou punição imposta na matéria deve ser proporcional ao fim legítimo buscado" (STF, 2015, p.62).

A Corte Interamericana, porém, continua a Ministra, "considerou conteúdo ao direito à liberdade de expressão e de pensamento, por ser formado tanto pelo direito de expressar o próprio pensamento (dimensão individual), quanto pelo direito de buscar, receber e difundir informações sobre qualquer matéria (dimensão social)".

4.2 Direito à informação (art. 5º, XIV)

O direito à informação é também é considerado constitucionalmente fundamental e se refere à proteção a se obter e divulgar informação sobre dados, qualidades, fatos de interesse da coletividade, ainda que sejam assuntos particulares, porém com expressão ou de efeitos coletivos.

Da liberdade de informação depende a opinião publica que se baseiam, exatamente, em dados que precisam ser públicos sobre pessoas que exercem funções públicas, por exemplo, no entanto, citando Agostini, a Ministra entende que na sociedade atual, apresenta-se diluída a "fronteira entre privado e político, ao ponto de se tornarem quase irreconhecíveis" (STF 2015, p. 74).

Citando Canotilho, a Ministra Cármem Lúcia, defende que em respeito à intimidade existem instrumentos jurídicos privilegiados de garantia tais como o direito à inviolabilidade do domicílio e da correspondência, da proibição de tratamento informático de dados referentes à vida privada, bem como o sigilo profissional, reserva confidencial para rostos e papéis pessoais e que é recente a constitucionalização expressa da inviolabilidade da intimidade, privacidade, honra e imagem (STF, 2015).

A proteção aos direitos à intimidade e privacidade foram deixados a cargo do legislador infraconstitucional: Código Penal Brasileiro, nos artigos 138 a 145, que versam sobre os crimes contra a honra; Lei nº 12.527/2011, Lei de Acesso à Informação (STF, 2015) e a Lei Geral de Proteção de Dados Pessoais, Lei nº 13.709/18, a qual foi aprovada recentemente, em 14 de agosto, 2018, dentre outras.

Em seu voto, a Ministra faz duas observações importantes sobre a constitucionalidade das normas questionadas nesta ação: (1) Definição de seu conteúdo na forma protegida no sistema jurídico fundamental brasileiro e de sua eficácia social; (2) A definição desse direito não é a mesma para todos, mas altera-se segundo a escolha feita pelo sujeito de direito a submeter-se a atividade que: a) seja regida pela transparência referente às instituições estatais, visando o controle que os cidadãos exercem sobre o governo; b) aqueles cuja condição profissional dependa do público para o seu desempenho com êxito; c) extraia ou retire dos cidadãos, pelo exercício de sua função ou atividade, os ganhos materiais, profissionais ou de reconhecimento, pelo que há de ser por eles conhecido.

Resumindo suas proposições, a Ministra afirma que em certos casos a pessoa se sujeita a ser conhecida e reconhecida pelo público. Assim, não se pode, portanto, considerar da mesma forma alguém que nada quer ou se extrai do público em sua condução de vida.

Especificamente em relação à questão da publicação das biografias, após tecer vários comentários sobre a importância histórica e social das mesmas, a Ministra afirma que a notoriedade exige o pagamento de um preço que "é fixado pela extensão da fama", que normalmente é buscada. Mesmo quando não é buscada a fama cobra seu preço que é o "reconhecimento público que se traduz em exposição do espaço particular", no qual todos querem adentrar (STF, 2015, p.87). No entendimento da ministra deve ser defendida a vida do outro, mas a curiosidade das pessoas também precisa ser satisfeita e para isso existe o trabalho do biógrafo (STF, 2015, p.99).

De qualquer forma, garante Cármem Lúcia, há sempre um espaço de indevassabilidade e segredo que compõe o íntimo de uma pessoa que não é pra ser levado ao conhecimento do público.

O passado compõe o que a pessoa se torna, a interpretação plural, advinda de variadas biografias é que podem levar gerações futuras a chegar à conclusão sobre o que ocorreu, porque e como se repetir (se positivo) ou evitar (episódios negativos).

No âmbito da proteção de dados, a exploração de dados pessoais ficará limitada por três princípios norteadores: princípio da finalidade (aquela que foi informada ao titular das informações); o princípio da minimização da coleta (só poderão ser coletados os dados realmente necessários para que seja atingida a finalidade); e, por fim, o princípio da retenção mínima (os dados deverão ser excluídos após atingirem sua finalidade).

Tendo em vista, porém, que foi vetado o artigo que criava uma agência específica para fiscalização da aplicação da lei (Autoridade Fiscalizadora), Schreiber (2018), entende serem equivocadas as propostas no sentido de atribuir esse poder de fiscalização para órgãos de segurança. O autor destaca ainda que a criação de uma Autoridade de Proteção de Dados seria fundamental para que a lei cumprisse seu papel efetivamente, a fim de criar uma cultura de privacidade.

5 CONSIDERAÇÕES FINAIS

Por ser uma decisão unânime na Suprema Corte brasileira, sobre uma ação direta de inconstitucionalidade dos artigos 20 e 21 do Código Civil pátrio, a relevância jurídica traz importante

impacto na jurisprudência e doutrina que versam sobre o tema no Brasil, as quais precisam se ajustar coerentemente, pela necessária estabilidade jurídica, conforme dispõe o Código de Processo Civil de 2015, em seu artigo 926.

Nos dias atuais, as facilidades das tecnologias da informação e da comunicação fazem com que determinados grupos de pessoas encontrem vantagens no relacionamento com o público em geral.

Mesmo que "a notoriedade torne a pessoa alvo de interesse público", as informações que extrapolem as linhas da quadra de jogo ou desempenho, a busca, produção e divulgação de informações não é ilegítima, "nem pode ser cerceada sob o argumento de blindar-se a pessoa com a inviolabilidade constitucionalmente assegurada" (STF, 2015, p.94).

Por um lado, as pessoas se beneficiam de câmeras para sua própria segurança; por outro, essa câmera também pode ser responsável pela captação de outras imagens que não deveriam, mas são divulgadas para o público em geral em questão de segundos.

Por todo o exposto, a Ministra destacou que é preciso saber se preservar, e que pessoas podem vir a serem sujeitos de estudos, pesquisas, obras, nas quais suas vidas são relatadas e os escritos produzidos e divulgados, independente de autorização da narrativa e das versões do autor da produção.

Assim, o estudo traz como conclusão, a partir da apresentação dos argumentos utilizados pela Ministra Cármem Lúcia, que o panoptismo idealizado por Bentham e destacado por Foucault, em seus estudos sobre a disciplina e o controle social, não é, nos dias atuais, um mecanismo de poder que beneficia apenas atividades de vigilância do aparelho estatal; Há interesses corporativos ou comerciais (lucro) e pessoais que mitigam o direito constitucional à privacidade, dessa forma, com a promulgação da Lei Geral de Proteção de Dados Pessoais, que entrará em vigor somente em fevereiro de 2020, os usuários poderão ter um controle maior sobre esse tipo de processamento de seus dados pessoais, com finalidade de lucro.

Assim, atinge-se o objetivo deste trabalho constatando-se que a tendência jurídica e jurisprudencial pelo que se depreende dos argumentos expressos no julgamento da Ação Direta de Inconstitucionalidade, ora estudada, é a de que o direito de privacidade e de intimidade seja considerado menos importante que os direitos de liberdade de expressão e liberdade de conhecimento.

Por último, cabe ressaltar que muito se discutirá ainda sobre as disposições da Lei Geral de Proteção de Dados Pessoais, até que esta entre em vigor (fevereiro/2018). Assim, além do estabelecimento de uma agência fiscalizadora para a proteção dos dados pessoais, também seria útil que outras pesquisas acompanhassem os desdobramentos do Regulamento Geral de Proteção de Dados da Europa, que já entrou em vigor em maio de 2018, e que exerceu forte influência na elaboração da lei brasileira que versa sobre o mesmo tema.

REFERÊNCIAS

BAUMAN, Zygmunt. **Vigilância Líquida**. Rio de Janeiro: Zahar, 2014.

BENTHAM, Jeremy. **O Panóptico**. 2ª ed. Belo Horizonte: Autêntica, 2008.

BRASIL. **Código de Processo Civil**. Senado Federal: Centro Gráfico, 2015.

_____. **Código Civil**. Senado Federal: Centro Gráfico, 2015.

_____. Constituição (1988). **Constituição da República Federativa do Brasil.** Senado Federal: Centro Gráfico, 1988.

_____. Supremo Tribunal Federal, **Ação Direta de Inconstitucionalidade** nº 4.815. Relatora Ministra Cármem Lúcia Antunes Rocha, j. 10-6-2015, P, DJE de 1º-2-2016.

CANCELIER. Mikhail Vieira de Lorenzi. O Direito à Privacidade hoje: perspectiva histórica e o cenário brasileiro. **Revista Sequência (Florianópolis)**, n. 76, p. 213-240, ago. 2017.

CASTELLS, Manuel. **A sociedade em rede.** São Paulo: Paz e Terra, 2003.

FOUCAULT, Michel. **Vigiar e Punir: nascimento da prisão.** Petrópolis: Vozes, 1987.

JABUR, Gilberto Haddad. **Liberdade de pensamento e direito à vida privada: conflito entre direitos da personalidade**. São Paulo: RT, 2000.

LAFER, Celso. **A reconstrução dos direitos humanos**. São Paulo: Companhia das Letras, 2006.

SCHEIBER, Anderson. **Proteção de Dados Pessoais no Brasil e na Europa**. Jornal Carta Forense, Colunas, Ensaio, 2018. Disponível em:<http://www.cartaforense.com.br/conteudo/colunas/protecao-de-dados-pessoais-no-brasil-e-na-europa/18269>. Acesso em 19 nov. 2018.

VIEIRA, Tatiana Malta. **O direito à privacidade na sociedade da informação: efetividade desse direito fundamental diante dos avanços da tecnologia da informação**. Dissertação de Mestrado no curso de pós-graduação stricto sensu em Direito, Estado e Sociedade: Políticas Públicas e Democracia. Universidade de Brasília, 2007.

WEIMANN, Gabriel. **Terror on the Internet: The New Arena**, the New Challenges, Special Report, United States Institute of Peace, Washington-DC, 2004.

EVASÃO ESTUDANTIL E CIÊNCIA DE DADOS:
primeiros passos de uma pesquisa aplicada no contexto da Educação a Distância da Universidade Estadual do Centro-Oeste

STUDENT EVASION AND DATA SCIENCE:
first steps of an applied research to the Distance Learning context of the Universidade Estadual do Centro-Oeste

Sandro Rautenberg[1], Alan H. Costa[1], Paulo R. V. do Carmo[1], Renan A. M. Nutse[1], Maria A. C. Knüppel[1], Marta C. R. Anciutti[1]
(1) Universidade Estadual do Centro-Oeste, Alameda Élio Antonio Dalla Vecchia, 838,
Vila Carli, Guarapuava-PR, CEP 85040-167,
{sandro.rautenberg, alanhenschel2, pauloviviurka4, renanmnutse, knuppelc, martanciutti}@gmail.com

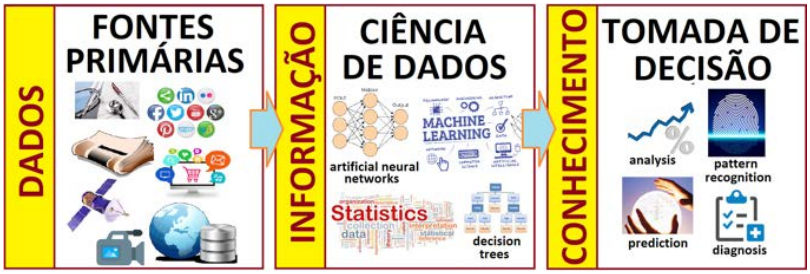
Resumo:
A Educação a Distância é uma alternativa de ensino para indivíduos que necessitam horários e locais de estudo flexibilizados. Dentre as instituições que oferecem essa modalidade de ensino, a Universidade Estadual do Centro-Oeste (UNICENTRO) ministra cursos de graduação a distância por intermédio do Núcleo de Educação a Distância (NEaD). Em uma análise realizada, o NEaD destaca um alto índice de evasão de alunos. Para melhor entender essa realidade, institucionalmente, formalizou-se um projeto de extensão empregando tecnologias da Ciência de Dados. Iniciado em Junho/2018, o projeto visa inferir alguns elementos de análise e apoiar decisões pedagógicas a partir dos microdados do Ambiente Virtual de Aprendizagem MOODLE. Neste sentido, este artigo tem como objetivo relatar os primeiros passos do referido empreendimento. Decorridos seis meses de execução, dentre os resultados preliminares alcançados, pontua-se: i) o estabelecimento de um *workflow* para desenvolver os cenários de exploração e visualização de dados; ii) a concepção de uma camada para ingestão de dados, mitigando os efeitos indesejáveis de versionamento de modelos de dados do MOODLE; e iii) a formalização de um meio de comunicação para com os gestores do NEaD/UNICENTRO, quando da prospecção de cenários de exploração de dados.
Palavras-chave: Ciência de Dados. Educação à Distância. Evasão Estudantil. MOODLE. Ambientes Virtuais de Aprendizagem.

Abstract:
Distance Learning is a teaching alternative for individuals who need flexible schedules and places for studying. Among the institutions that offer this kind of education, the Universidade Estadual do Centro-Oeste (UNICENTRO) teaches undergraduate distance courses through the Núcleo de Educação a Distância (NEaD). In an initial analysis, the NEaD highlights a high evasion rate among students. To better understand this reality, institutionally, a project was formalized using Data Science technologies. Initiated in June/2018, the project aims to infer some elements from the microdata of the MOODLE Learning Management System for supporting pedagogical decisions. In this sense, this article aims to report the first steps of this effort. After six months of execution, the preliminary results achieved include: i) the establishment of a workflow for developing data exploitation scenarios; ii) the design of a data layer, mitigating the undesirable effects of versioning of the MOODLE's data model; and iii) the formalization of a communication model with the NEaD/UNICENTRO managers, when prospecting data exploitation scenarios.
Keywords: Data Science. Distance Learning. Student Evasion. Moodle. Learning Management System.

1 INTRODUÇÃO

Com o avanço do uso da Internet, produz-se cada vez mais dados em várias plataformas digitais. Diversos aplicativos e dispositivos interconectados (computadores, *smartphones*, etc.) relacionam uma série de eventos (van der AALST, 2014), armazenando enormes quantidades de registros, sinais, imagens, vídeos e *posts*. Por conseguinte, os dados são abundante e velozmente produzidos, podendo servir como matéria-prima à Tomada de Decisão (ECONOMIST, 2018). Por isso, o desenvolvimento de soluções computacionais que obtém insumos de conhecimento de volumes de dados torna-se foco de investimento das organizações.

Figura 1. Contexto do Trabalho.



Fonte: Dados da Pesquisa.

Atualmente, esse contexto relaciona os conceitos de Ciência de Dados e Tomada de Decisão como segue. Conforme ilustrado na Figura 1, tem-se as fontes de dados cujas características dificultam a captura, o armazenamento, o gerenciamento, a análise e a exploração de dados por parte de ferramentas computacionais tradicionais (GARTNER, 2018; MANYIKA *et al.*, 2011). Para auxiliar a análise e a exploração de dados, pode-se recorrer à Ciência de Dados (do inglês, *Data Science*). A Ciência de Dados é caracterizada como uma camada de métodos devotados à extração de informação útil a partir de complexas e dinâmicas bases de dados (BUGNION; MANIVANNAN; NICOLAS, 2017). E, por conseguinte, ao recuperar informação útil, pode-se auxiliar os gestores no desempenho de suas atividades decisórias.

Diante dessa visão, este trabalho parte do pressuposto que as organizações voltadas ao Ensino a Distância (EaD)[1] podem se beneficiar dos métodos e tecnologias da Ciência de Dados para melhorar seus processos. Principalmente, em face destas organizações adotarem Tecnologias da Informação e Comunicação (TICs), produzindo conjuntos de dados a serem explorados, a partir dos: **i)** gerenciamento da vida acadêmica de seus discentes; e **ii)** uso de meios digitais para o compartilhamento de material educacional. No tocante às referidas organizações, um dos problemas principais enfrentados é a Evasão Estudantil[2], fenômeno cuja taxa média anual alcança 25% dos alunos matriculados (ABED, 2014).

1 "[...] a EaD pode ser entendida como agente de inovação dos processos de ensino e aprendizagem, incentivando a incorporação de novas Tecnologias da Informação e Comunicação (TICs) aos métodos didático-pedagógicos, e possibilitando ao cidadão o acesso à educação superior pública e de qualidade, a partir da democratização do acesso à educação [...]" (MORÉ *et al.*, 2010).

2 A Evasão Estudantil é compreendida como "a saída definitiva do aluno de seu curso de origem, sem concluí-lo" (ANDIFES; ABRUEM; SESu/MEC, 1996).

Pontualmente, esse problema também é enfrentado pelo Núcleo de Educação à Distância da Universidade Estadual do Centro-Oeste (NEaD/UNICENTRO). O NEaD/UNICENTRO tem como incumbência a difusão do conhecimento tecnológico e científico para os diferentes segmentos sociais, por intermédio de projetos, atividades e programas de EaD. Mediante seus polos avançados, essa organização é capilarizada em 51 municípios no Estado do Paraná (NEAD, 2018), circunvizinhando quase todos os 399 municípios da referida unidade federativa. Institucionalizado no ano de 2005, atualmente, o NEaD/UNICENTRO enfrenta o desafio de minimizar seus índices de Evasão Estudantil, que está em torno de 54%.

Tecnologicamente, para minimizar os índices de Evasão Estudantil, Silva (2017) pontua a construção de modelos preditivos. Tais modelos deveriam identificar preventivamente a probabilidade de um aluno evadir. Segundo a autora, com a utilização de modelos preditivos, pode-se abstrair informações confiáveis para suporte ao processo de Tomada de Decisão dos gestores, aumentando os índices de retenção discente. Portanto, admite-se que o emprego de métodos e tecnologias da Ciência de Dados é pertinente ao estabelecimento de modelos preditivos da Evasão Estudantil.

Diante disso, para melhor entender e tratar a realidade da Evasão Estudantil inerente ao NEaD/UNICENTRO, um Projeto de Extensão[3] foi formalizado em Junho de 2018. Em sua essência, tal empreendimento visa criar modelos de análise e visualização de dados e informações da evasão, baseando-se em metodologias e tecnologias da Ciência de Dados. Cabe ressaltar que objetivos do referido projeto permeiam, principalmente: **i)** a abstração dos motivos da desistência dos alunos matriculados em cursos de graduação na modalidade EaD; e **ii)** o amparo em ações afirmativas para aumentar os índices de retenção de estudantes.

2 OBJETIVO

Diante o exposto, o objetivo deste capítulo de livro é relatar os resultados parciais alcançados no tocante aos primeiros meses de atividades realizadas do projeto de extensão institucionalizado.

3 MATERIAIS E MÉTODOS

Para subsidiar as atividades de EaD, o NEaD/UNICENTRO utiliza o Ambiente Virtual de Aprendizagem MOODLE (*Modular Object Oriented Distance LEarning*) como plataforma de suporte (versão 3.x). Este é um sistema gratuito que permite a gestão de cursos *online*, mediante o gerenciamento da comunicação entre atores (docentes, discentes, tutores, etc.) e dos objetos de aprendizagem (*ebooks*, vídeo aulas, atividades, fóruns, mensagens, etc.). Assim, o MOODLE oferece um ambiente dinâmico que oportuniza o aprendizado a qualquer momento e em qualquer lugar, ao utilizar a Internet como plataforma de comunicação (MOODLE, 2018).

3 RESOLUÇÃO Nº 022 - PROEC/UNICENTRO, DE 16 DE JULHO DE 2018 - Aprova o Projeto de Extensão Data Science e evasão estudantil: um estudo do caso no contexto do Núcleo de Educação a Distância da Uni-versidade Estadual do Centro-Oeste, na modalidade de Ação de Extensão, na categoria de Projeto de Extensão (UNIVERSIDADE ESTADUAL DO CENTRO-OESTE - UNICENTRO, 2018).

Atualmente, a base de dados do MOODLE do NEaD/UNICENTRO comporta cerca de 16 *gigabytes* de dados, contendo os registros de 5.785 alunos de EaD (formados, desistentes ou em formação). De acordo com seu modelo relacional, na referida base de dados, os registros da interação dos usuários para com os objetos disponibilizados são armazenados minuciosamente em suas 300 tabelas de dados. Diante dessa riqueza de detalhes em seus registros, a base de dados do MOODLE é um insumo pertinente ao discernimento da Evasão Estudantil inerente ao NEaD/UNICENTRO.

Neste sentido, para promover a análise e a exploração de dados/informação a partir da base de dados do MOODLE, adota-se a proposta de Bugnion; Manivannan e Nicolas (2017) como procedimento metodológico da Ciência de Dados. Os referidos autores sugerem sete passos (Figura 2), conforme segue:

- **Obtenção de Dados**. Preconiza as tarefas de avaliação e seleção de dados primários e seus metadados, por exemplo, a partir: do processamento de arquivos de texto; do monitoramento de uma rede de sensores; de consultas a bases de dados de sistemas legados; ou de API (*Application Programming Interface*) da web.
- **Ingestão de Dados**. Trata das transformação e carga dos dados primários advindos de fontes diferentes e formatos diversificados em uma base de dados centralizada. Isso implica em organizar, representar e inserir dados pré-processados em um repositório de dados principal, mitigando os esforços futuros na geração de informação relevante.

Figura 2. Ciclo de Vida da Ciência de Dados.

Fonte: baseado em (BUGNION; MANIVANNAN; NICOLAS, 2017) [tradução dos autores]

Exploração de Dados. Privilegia a execução de estudos preliminares de modo a estabelecer conjecturas iniciais e entendimentos superficiais dos dados disponibilizados em relação à informação requisitada. Assim, estabelece-se um fluxo de trabalho (*workflow*) para relacionar os dados em busca da informação relevante.

Definição dos Parâmetros. Está intimamente ligado às escolhas necessárias para o emprego do(s) algoritmo(s) de Aprendizado de Máquina. Nesta atividade, por exemplo: converte-se os dados de

entrada conforme os requisitos de manipulação do algoritmo de aprendizado; transforma-se os dados de saída de modo a refletir uma saída legível aos seres humanos; estabelece-se os intervalos dos parâmetros de entrada a serem considerados; define-se os critérios de parada do algoritmo de aprendizado ou o nível de confiabilidade exigido da resposta gerada.

Implementação do Modelo. Prima pela utilização dos algoritmos de Aprendizado de Máquina para estabelecer modelos a partir dos dados de entrada e saída. Iterativamente, envolve o emprego de estratégias de treinamento e testes dos algoritmos para a definição dos parâmetros mais adequados. Como resultado, um modelo é estabelecido que melhor represente as características dos dados utilizados.

Utilização do Modelo. Uma vez estabelecido um modelo, pode-se utilizá-lo para inferir informações sobre dados não utilizados na etapa anterior. Isso confirmará o poder de generalização do modelo perante situações do mundo real.

Tomada de Decisão. Em situações reais, combinando o resultado gerado pelo modelo com o conhecimento particular do domínio, os gestores tomam suas decisões. Uma parte fundamental nesta etapa envolve a customização da apresentação/visualização dos dados e informações através de relatórios e gráficos. Desta forma, é possível gerar *insights* mais claros e convincentes em um processo de Tomada de Decisão.

Como suporte tecnológico, salienta-se que no Projeto de Extensão é previsto o uso da Linguagem de Programação Python. No contexto do desenvolvimento de softwares, Python destaca-se por ser uma linguagem de programação de alto nível, interpretada, baseada no paradigma de orientação a objetos e com dinamicidade semântica (PYTHON.ORG, 2018). Por dinamicidade semântica entende-se o poder de extensão da linguagem com a adição de novas funcionalidades mediante a incorporação de bibliotecas. Neste sentido, a este projeto poderão ser incorporadas as bibliotecas:
Pandas[4]. É uma biblioteca *open source* de alta performance que facilita o uso de estruturas e ferramentas de análise de dados em soluções computacionais desenvolvidas em Python.

Psycopg2[5]. Em aplicações desenvolvidas com Python, permite o acesso rápido e seguro a bases de dados mantidas no Sistema Gerenciador de Banco de Dados PostgreSQL.

Matplotlib[6]. É uma biblioteca para desenvolvimento de gráficos estatísticos, podendo ser utilizada em soluções computacionais Python.

Seaborn[7]. Assim como a Matplotlib, a Seaborn também permite o desenvolvimento de visualizações estatísticas de dados em aplicações Python.

4 Acesso: https://pandas.pydata.org/.
5 Acesso: http://initd.org/psycopg/docs/.
6 Acesso: https://matplotlib.org/.
7 Acesso: https://seaborn.pydata.org/.

NLTK (*Natural Language ToolKit*)[8]. É uma biblioteca feita para processamento de texto em linguagem natural. Possui ferramentas estatísticas e gráficas e *corpus* textuais para testes em vários idiomas.

VADER (*Valence Aware Dictionary and sEntiment Reasoner*)[9]. Juntamente com a NLTK, provê uma ferramenta léxica de suporte à análise de sentimentos. É baseada em regras sintáticas e semânticas, podendo ser utilizada na detecção de tendências positivas, neutras ou negativas em *post* de mídias sociais ou mensagens entre usuários, por exemplo.

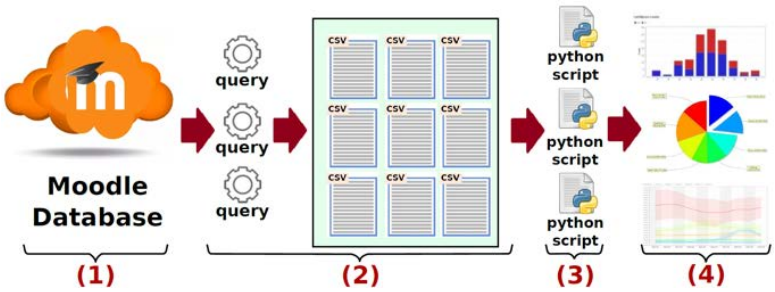
Scikit-learn[10]. É uma biblioteca *open source* que implementa vários algoritmos de Aprendizado de Máquina em Python. Inclui algoritmos de classificação, regressão e agrupamento, por exemplo.

4 DISCUSSÃO E RESULTADOS PARCIAIS

Considerando o ciclo de vida da Ciência de Dados, o estágio atual do projeto institucionalizado envolve a Ingestão de Dados e a Exploração de Dados. Ressalta-se que estes passos são relacionados aos dados e metadados da base de dados do MOODLE, dispendendo esforço significativo por parte dos cientistas de dados quanto ao: **i)** entendimento da complexidade do modelo de dados; e **ii)** pré-processamento e limpeza dos dados.

A Figura 3 representa as atividades desempenhadas junto aos gestores do NEaD/ UNICENTRO. As atividades envolvem:

Figura 3. Processo de Desenvolvimento de Cenários.



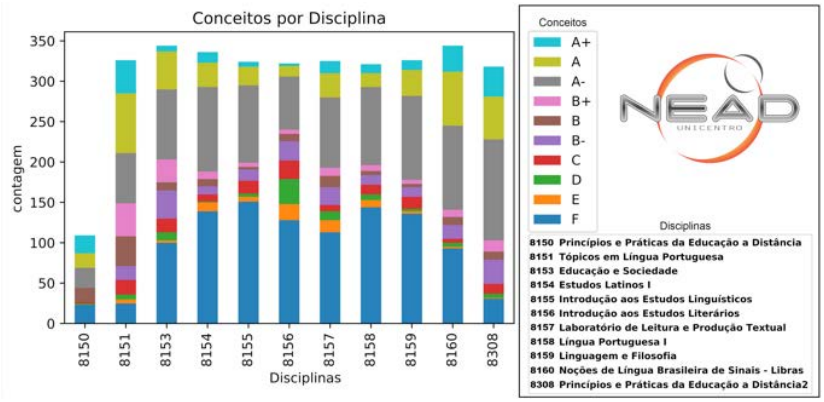
Fonte: Dados da Pesquisa.

1) A Obtenção de Dados é realizada junto à Coordenadoria de Tecnologia e Informação da UNICENTRO. A base de dados disponibilizada advém de um arquivo de *backup* do Ambiente Virtual de Aprendizagem MOODLE. Os dados obtidos são estagiados em uma instância local do Sistema Gerenciador de Banco de Dados PostGreSQL.

8 Acesso: https://www.nltk.org/.
9 Acesso: https://github.com/cjhutto/vaderSentiment.
10 Acesso: http://scikit-learn.org.

2) A Ingestão de Dados é realizada de forma incremental, à medida que os cenários de exploração ou visualização de dados/informação e os respectivos *workflows* são definidos em conjunto aos gestores do NEaD/UNICENTRO. A cada cenário definido, uma *query* é desenvolvida para organizar, coletar e armazenar um subconjunto de dados relevante no repositório centralizado.
3) Para cada subconjunto e cenário estabelecido, um *script* Python é desenvolvido para: **i)** gerar uma apresentação de dados ou visualização informações através de relatórios e gráficos, respectivamente (fase de Exploração de Dados); ou **ii)** generalizar um modelo abstrato a partir dos dados, mediante o emprego de algoritmos de Aprendizagem de Máquina (fases de Definição de Parâmetros e Implementação de Modelo).
4) Como meio de comunicação para com os gestores do NEaD/UNICENTRO, gráficos e relatórios são apresentados, fomentando a geração de *insights* em um processo decisório (fases de Utilização do Modelo e Tomada de Decisão).

Figura 4. Gráfico para comunicação com os gestores. Cenário proposto para auxiliar a análise da distribuição dos conceitos alcançados, disciplina a disciplina, por alunos em formação - curso de graduação EaD "Letras Português e Literatura de Língua Portuguesa".



Fonte: Dados da Pesquisa.

Como prova de conceito das atividades estabelecidas, na forma de um ensaio elementar, um cenário de visualização de dados (Figura 4) foi proposto pelos gestores do NEaD/UNICENTRO. Para com os objetivos deste trabalho, o propósito deste ensaio de visualização foi: **i)** ratificar a forma de trabalho entre os cientistas de dados; e **ii)** sensibilizar os gestores quanto à forma de comunicação.

5 CONSIDERAÇÕES PARCIAIS

Este capítulo de livro está circunscrito a um projeto institucional, o qual objetiva a produção de elementos de análise e visualização voltados à compreensão do fenômeno da Evasão Estudantil nos cursos de graduação EaD do NEaD/UNICENTRO. Considerando a evolução dados → informação → conhecimento, o projeto institucionalizado utiliza os microdados do Ambiente Virtual de Aprendizagem MOODLE para abstrair informação útil sobre os padrões de desistência dos alunos matriculados. Consequentemente, com as informações abstraídas submetidas ao conhecimento de

domínio dos gestores, pode-se amparar as ações que contribuem ao aumento dos índices de retenção dos estudantes. Em face disso, baseando-se em tecnologias da Ciência de Dados, pretende-se subsidiar um processo profícuo de Tomada de Decisão guiada por Dados no ambiente organizacional do NEaD/UNICENTO.

Neste contexto, considerando o ciclo de vida da Ciência de Dados (Bugnion; Manivannan; Nicolas, 2017) e o período de execução, o projeto encontra-se nas fases de ingestão e exploração de dados. Cronologicamente, decorreram seis meses de atividade. Apesar da infância do projeto, alguns resultados preliminares são de grande relevância ao NEaD/UNICENTRO, sendo:

i. O estabelecimento do *workflow* padrão de desenvolvimento de cenários (Figura 3).
ii. A concepção de uma camada independente para obtenção e ingestão de dados (as *queries* do passo 2 – Figura 3). Como uma camada de tradução de dados, sua utilização isola a ingestão de dados com *queries*, mitigando os efeitos indesejáveis de versionamento do modelo de dados do MOODLE. Tal feita padroniza a utilização dos dados na visualização da informação ou na generalização de modelos.
iii. A formalização de um meio de comunicação para com os gestores do NEaD/UNICENTRO. Mediante isso, incrementalmente, a prospecção de novos cenários e a compreensão do fenômeno da Evasão Estudantil podem ser facilitadas.

Ademais, como trabalhos futuros para com o projeto institucionalizado, pretende-se avançar à atividade de Utilização de Modelos de Aprendizagem de Máquina. Neste sentido, é prevista a implementação de soluções baseadas em Agrupamento, Análise de Sentimentos e Modelos Preditivos. Consequentemente, almeja-se aprimorar a Tomada de Decisão nas ações pedagógicas dos gestores do NEaD/UNICENTRO.

Agradecimentos

À Secretaria da Ciência, Tecnologia e Ensino Superior (SETI/PR) pelo suporte financeiro (Projeto - Implementação da Universidade Virtual do Paraná – UVPR/SETI, Termo de Cooperação nº 145/2017, vinculado a unidade gestora do Fundo Paraná).

REFERÊNCIAS

ABED – Associação Brasileira de Educação a Distância. **Censo EAD Brasil 2014 - Relatório Analítico da Aprendizagem a Distância no Brasil**. Curitiba, 2015. Disponível em: <https://goo.gl/kbFxEJ>. Acesso em: nov. 2018.

ANDIFES, A.; ABRUEM, A.; SESu/MEC, S. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas: resumo do relatório apresentado a ANDIFES, ABRUEM e SESu/MEC pela Comissão Especial. **Avaliação - Revista Da Avaliação Da Educação Superior**, v. 1, n. 2, p. 55-65, 1996.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. Sebastopol: O'Reilly Media, Inc., 2009.

BÜCHNER, A. **Moodle 3 Administration**. 3ª ed. Birmingham: Packt Publishing, 2016.

BUGNION, P.; MANIVANNAN, A.; NICOLAS, P. R. **Scala: Guide for Data Science Professionals**. Birmingham: Packt Publishing, 2017.

ECONOMIST. **The world's most valuable resource is no longer oil, but data**. Disponível em: <https://goo.gl/AW4XsF>. Acesso em: 28 jul. 2018.

GARTNER. **What is Big Data? – Gartner IT Glossary – Big Data**. Disponível em: <https://goo.gl/GwQWLA>. Acesso em: 28 jul. 2018.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C. B.; HUNG, A. **Big data: The next frontier for innovation, competition, and productivity**. Disponível em: <https://goo.gl/Vg2G2U>. Acesso em: 28 jul. 2018.

MOODLE. Features – MoodleDocs. Disponível em: <https://docs.moodle.org/35/en/Features>. Acesso em: 06 set 2018.

MORÉ, R. P. O.; MORITZ, G. de O.; PEREIRA, M. F.; MELO, P. A. de. Modelo de Gestão para Educação a Distância: o Sistema de Acompanhamento ao Estudante – SAE. **RAI - Revista de Administração e Inovação**, v. 7, n. 2, p. 104-125, 2010.

EaD. **Núcleo de Ensino à Distância da Unicentro - PR » Polos**. Disponível em: <https://goo.gl/wPYJqe>. Acesso em: 05 set. 2018.

PYTHON.ORG. **What is Python? Executive Summary | Python.org**. Disponível em: <https://goo.gl/QYghos>. Acesso em: 06 set. 2018.

SILVA, F. C. da. **Gestão da Evasão na EaD:**

Modelo Estatístico Preditivo para os Cursos de Graduação a Distância da Universidade Federal de Santa Catarina. Florianópolis, 2017. 137 f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Sócio-Econômico. Programa de Pós-Graduação em Administração.

UNIVERSIDADE ESTADUAL DO CENTRO-OESTE (UNICENTRO). **RESOLUÇÃO Nº 022 - PROEC/UNICENTRO, DE 16 DE JULHO DE 2018**. Disponível em: <https://goo.gl/7VcZX4>. Acesso em: 06 set. 2018.

van der AALST, W. Data Scientist: The Engineer of the Future. In: Interoperability of Enterprises Systems and Applications Conference (I-ESA'2014), 2014, Albi-France, **Proceedings…** Heidelberg: Springer, 2014.

GESTÃO DE DADOS DA BIODIVERSIDADE:
aplicação do padrão de metadados Darwin Core

BIODIVERSITY DATA MANAGEMENT:
applying the Darwin Core metadata standard

Filipi Miranda Soares[1], Raíssa Yuri Hamanaka[2], Benildes C. Moreira dos Santos Maculan[3]
(1) Escola de Ciência da Informação, Universidade federal de Minas Gerais, filipivgp2011@gmail.com
(2) Escola de Ciência da Informação, Universidade federal de Minas Gerais, raissa0201@gmail.com
(3) Escola de Ciência da Informação, Universidade federal de Minas Gerais, benildes@gmail.com

Resumo:
No contexto da gestão de dados de pesquisa e da necessidade de padronização do armazenamento e tratamento dos mesmos para posterior reutilização, torna-se imprescindível o uso de metadados para a representação dos dados. Devido a importância do uso de metadados para a recuperação de documentos, o estudo objetivou analisar a aplicação do padrão de metadados Darwin Core (DwC) em um registro de ocorrência do Portal da Biodiversidade. Para avaliar a aplicação do DwC pelo Portal da Biodiversidade, foi feita uma correlação entre os metadados do DwC com os metadados do registro de ocorrência do repositório, buscando-se entender o significado de cada metadado e se eles correspondem de fato aos metadados do DwC, uma vez que a documentação do Portal da Biodiversidade apresenta o DwC como padrão de metadados adotado. Dos 34 campos do registro de ocorrência analisados, a maioria tinha correspondência com os campos do padrão de metadados. Houve casos de duplicação de campos no repositório da biodiversidade, da criação de campos no repositório que não existiam no DwC e da existência de campo inadequado no repositório. A partir do exemplo do Portal da Biodiversidade é possível delinear a importância da curadoria digital, ao procurar agrupar conjuntos de dados semelhantes e estruturá-los de forma a possibilitar o reuso dos mesmos.
Palavras-chave: Gestão de dados de pesquisa. Metadados. Darwin Core.

Abstract:
In the context of the management of research data and the need to standardize the storage and treatment of them for later reuse, it is essential to use metadata to represent the data. Due to the importance of using metadata for document retrieval, the study aimed to analyze the application of the Darwin Core (DwC) metadata standard in a Portal da Biodiversidade occurrence record. To evaluate the application of the DwC by the Portal da Biodiversidade, a correlation was made between the metadata of the DwC with the metadata of the record of occurrence of the repository, trying to understand the meaning of each metadata and if they correspond in fact to the metadata of the DwC, since the documentation of the Portal da Biodiversidade presents DwC as the metadata standard adopted. Of the 34 occurrence record fields analyzed, most corresponded to the metadata standard fields. There have been cases of duplication of fields in the biodiversity repository, the creation of fields in the repository that did not exist in the DwC and the existence of an inadequate field in the repository. Based on the example of the Biodiversity Portal, it is possible to delineate the importance of digital curatorship by seeking to group similar data sets and structure them in a way that allows them to be reused.
Keywords: Research data management. RDM. Metadata. Darwin Core.

1 INTRODUÇÃO

No atual contexto do fenômeno *big data*, em que são gerados grandes volumes de dados não-estruturados em meio digital, torna-se de extrema importância a questão da organização do dado para sua posterior recuperação.

Inseridos nesse contexto estão os pesquisadores, produzindo um volume grande de dados científicos, o que impacta na definição das atuais fontes de informação (para além do periódico científico) e no compartilhamento de dados, e, em consequência, na comunicação científica. Dessa forma, surgem desafios na gestão dos dados de pesquisa disponíveis em rede (SAYÃO; SALES, 2012).

Como solução para esses desafios tem origem o conceito de curadoria digital que "além de reduzir a duplicação de esforços na criação de dados de pesquisa [...] reforça o valor de longo prazo dos dados existentes quando os tornam disponíveis para a reutilização em novas pesquisas" (SAYÃO; SALES, 2012, p. 184).

Para garantir o reuso dos dados de pesquisa por pesquisadores, é necessário que os sistemas computacionais consigam acessá-los. Assim, tanto dados quanto algoritmos devem ser criados segundo os princípios de reusabilidade, encontrabilidade, acessibilidade e interoperabilidade (FAIR, 2016). Isso pode ser problematizado pelas seguintes questões: "se você disponibilizar seus dados para um cientista ou pesquisador que não esteve envolvido com seu projeto, seriam capazes de entendê-lo?"; "eles conseguiriam utilizar os dados efetiva e apropriadamente?"; "como criar, organizar, gerir, descrever, preservar e compartilhar dados, efetivamente?" (STRASSER et al., 2012, p. 1, tradução nossa).

A efetiva gestão de dados de pesquisa e sua posterior reutilização por demais pesquisadores, depende de um conjunto de práticas durante a coleta, o processamento e a análise dos mesmos. Essas práticas compõem o ciclo de vida dos dados (CVD), que é formado pelos seguintes elementos: planejar, coletar, assegurar, descrever, preservar, descobrir, integrar e analisar (STRASSER et al., 2012). Para Sant'ana (2016), cada etapa ou elemento do CVD possui características em comum: privacidade, integração, qualidade, direito autoral, disseminação e preservação. Sayão e Sales (2012) dividem o CVD em ações: para todo o ciclo de vida (descrição e representação da informação, planejamento da preservação, participação e monitoramento e curadoria e preservação); sequenciais, que se repetem em cada etapa do ciclo (conceitualização, criação ou recebimento, avaliação e seleção, arquivamento, preservação, armazenamento, acesso, uso e reuso e transformação) e ocasionais (eliminação, reavaliação e migração).

O foco deste estudo foi o elemento "descrição" do CVD, definido como "dados que são descritos com precisão e minuciosamente usando os padrões de metadados apropriados" (STRASSER et al., 2012, p. 3, tradução nossa).

Conforme Costa (2017), há algumas iniciativas brasileiras em relação à gestão de dados de pesquisa, e dentre elas se destacam: o Programa FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) de Pesquisa em e-Science; o Portal da Biodiversidade; a Infraestrutura Nacional de Dados Espaciais no Brasil (INDE); as medidas de incentivo ao acesso aberto do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e o desenvolvimento do repositório de dados do Instituto de Energia Nuclear.

Para explicitar o elemento descrição (e seus fatores) do CVD, analisaram-se os metadados de um registro de ocorrência do repositório de dados do Portal da Biodiversidade em comparação com o padrão de metadados Darwin Core (DwC). DwC é um padrão internacional com base em *taxon*, composto por um conjunto de elementos (propriedades, atributos ou conceitos) desenvolvidos para padronizar o compartilhamento de informações sobre diversidade biológica.

2 Objetivo

Analisar a aplicação do padrão de metadados Darwin Core (DwC) em um registro de ocorrência do Portal da Biodiversidade[1].

3 PROCEDIMENTOS METODOLÓGICOS

O universo de pesquisa é todos os registros no Portal da Biodiversidade e como amostra foi selecionado o registro de ocorrência da espécie *Rhinella granulosa* (Spix, 1824), leigamente conhecida como sapo da caatinga. O procedimento adotado foi correlacionar os metadados do DwC com o significado atribuído a cada campo do registro e avaliar o seu uso. Assim, as categorias de análise selecionadas para análise foram os metadados adotados para descrição do registro de ocorrência, que foram correlacionadas com o conjunto de metadados do DwC.

4 RESULTADOS

A análise determinou que o registro selecionado possui 34 metadados, apresentados em formato de folha de dados (CSV[2]), baixado do Portal da Biodiversidade, que serão apresentados na ordem que aparecem no registro.

Os dois primeiros metadados do registro são <Nome da instituição> e <Sigla da instituição>. Eles representam um mesmo dado e possuem o mesmo equivalente no DwC, que é o metadado <institutionCode>. O primeiro campo de metadado está sem informações e o segundo apresenta a sigla ICMbio. Esses dois campos deveriam ser apenas um campo, pois significam a mesma coisa. Assim, seria necessário apenas definir o formato de entrada de dados: nome por extenso da instituição ou apenas a sigla.

Em seguida, aparece o metadado <Nome da base de dados>, que possui equivalente no DwC como <datasetName> e está preenchido corretamente. O campo seguinte foi denominado <Sigla da base de dados> e não possui equivalente no DwC. A melhor prática sugerida seria unir esses dois campos em apenas um e padronizar a entrada de dados: nome por extenso da base de dados ou apenas sigla.

Depois vem o campo <Responsável pelo registro> que possui o metadado equivalente no DwC <recordedBy>, com valor formatado de acordo com as recomendações do DwC para este campo.

1 Disponível em: <https://portaldabiodiversidade.icmbio.gov.br/portal>. Acesso em: 30 set. 2018.

2 *Comma-separated values* é um formato de arquivo que apresenta dados tabelados.

Em sequência, são apresentados dois campos que equivalem a um mesmo campo no DwC, mas foram preenchidos com valores distintos: <Número do registro no portal> e <Número do registro na base de dados>, que equivalem semanticamente ao campo <catalogNumber> no DwC. O primeiro apresenta o valor numérico '1067665' e o segundo o valor "Nº Da Autorização/Licença Sisbio: 44832".

Após, aparece o campo <Data do registro> que é ambíguo. Não é possível saber se o valor da data que preenche o campo é a data de registro na base de dados ou a data que o evento (ocorrência) aconteceu. No padrão DwC, o metadado para descrever a data de ocorrência é o <eventDate>, presente no registro como <Data do evento>. Quando baixado em formato CSV, o valor do campo <Data do registro> muda de '05/02/2017' para '06/02/2017', a mesma data presente no campo <Data do evento>. Não há um campo de data de registro no DwC. O campo com significado mais próximo seria <dcterms:modified>, que se refere à última data de modificação do registro documental da ocorrência na base de dados. Além disso, o DwC recomenda utilizar esquemas de codificação para preenchimento das datas, como a ISO 8601:2004(E), o que não foi aplicado no Portal da Biodiversidade. O valor do campo <Data do evento> apresenta outro problema: está definido como '06/02/2017 a 06/02/2017'. Se o evento aconteceu no dia seis de fevereiro e não foi inserido o intervalo específico de horas em que ele ocorreu, não é necessário inserir a data como intervalo de datas, uma vez que o evento aconteceu em apenas um dia.

O metadado seguinte, <Data de Carência>, não possui equivalente no DwC e não está claro ao que esta data se refere.

Depois, aparece o metadado <Nome científico>, que possui o equivalente <scientificName> no DwC. O valor desse campo no registro é 'Rhinella granulosa', que não está de acordo com as recomendações do DwC (DARWIN CORE TASK GROUP, 2015), que orienta indicar o nome científico completo, com autor e data. A melhor prática seria apresentar o nome '*Rhinella granulosa* (Spix, 1824)'.

Em seguida vem <nome comum>, que possui o equivalente <vernacularName> no DwC, e está preenchido em conformidade com as melhores práticas recomendadas pelo padrão.

O campo seguinte, <Nome científico na base de dados>, não deveria existir. O nome científico da espécie é um identificador único, ou seja, não pode variar. Logo, é desnecessário que haja dois campos com o mesmo valor no registro de ocorrência.

Logo em seguida aparecem os metadados <Nível taxonômico>, <Número de indivíduos>, <Reino>, <Filo>, <Classe>, <Ordem>, <Família> e <Gênero>, que têm valores que atendem às orientações de melhores práticas do DwC, e possuem como equivalentes, respectivamente: <taxonRank>, <individualCount>, <kingdom>, <phylum>, <class>, <order>, <family> e <genus>.

Depois vem o metadado <Espécie>, que repete o mesmo valor de dois outros campos do registro de ocorrência, portanto, é redundante sua presença no registro. A sugestão seria substituir esse metadado por <specificEpithet>, que representa o epíteto[3] da espécie no nome científico.

3 Designa uma espécie ou subespécie diferente dentro de um mesmo gênero.

A seguir aparecem os metadados <Estado de conservação>, <Categoria de Ameaça> e <Status de Sensibilidade>, que ainda não têm equivalência no DwC. Entretanto, uma extensão[4] está sendo desenvolvida para representar informações de estado de conservação.

Após, aparecem metadados com informações de localização: <Localidade>, <País>, <Estado/Província>, <Município>, <Latitude> e <Longitude>, que possuem equivalentes no DwC, respectivamente: <locality>, <country>, <stateProvince>, <municipality>, <decimalLatitude> e <decimalLongitude>, e têm valores conforme as orientações do padrão.

Depois aparecem os campos <Outras informações da localidade> e <Jurisdição>, que não possuem equivalentes no DwC.

Por fim, o metadado <Destino do Material> tem o equivalente no DwC <MaterialSample>.

5 CONSIDERAÇÕES FINAIS

Conforme o objetivo do estudo, foram analisados 34 campos de um registro de ocorrência do Portal da Biodiversidade em correlação com os campos do padrão de metadados DwC. Na maioria das análises, em cerca de 65% dos casos, houve correspondência do campo do registro com o campo do padrão de metadados. Entretanto, houve casos de duplicação de campos no repositório da biodiversidade, valores inadequados, criação de campos que ainda não existem no DwC e existência de campo inadequado no repositório.

O Portal da Biodiversidade reuni registros de ocorrência de nove bases de dados de biodiversidade brasileira, sendo uma iniciativa de gestão de dados que desde 2016 está integrado ao Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr), que está ligado ao Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC).

Essa iniciativa é um exemplo que permite perceber a importância da curadoria digital, que procura agrupar conjuntos de dados semelhantes e estruturá-los de maneira a possibilitar o seu reuso. Para tanto, é necessário manter interoperabilidade entre as diferentes bases, o que pode ser alcançado por meio do uso de padrões de metadados. Considera-se que na implementação do CVD, os metadados são essenciais, pois possibilitam a padronização da representação e/ou codificação dos dados para sua futura recuperação.

Acredita-se que a padronização dos dados coletados e armazenados em repositórios permite a democratização dos dados ao garantir o acesso efetivo aos mesmos.

REFERÊNCIAS

COSTA, Maíra Murrieta. **Diretrizes para uma política de gestão de dados científicos no Brasil**. 2017. 288 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, DF, 2017.

4 Disponível em: <https://tools.gbif.org/dwca-validator/extension.do?id=http://purl.org/plic/terms/3.2.1/ThreatStatus>. Acesso em: 25 set. 2018.

DARWIN CORE TASK GROUP. Biodiversity Information Standards. Darwin Core. 2015. Disponível em: <http://rs.tdwg.org/dwc/>. Acesso em: 8 Set. 2018.

FAIR principles for data stewardship. **Nature Genetics**, London, v. 48, n. 4, p. 343, Apr. 2016.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. **Inf. Inf.**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016.

SAYÃO, Luis Fernando; SALES, Luana Faria. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Inf. & Soc.: Est.**, João Pessoa, v. 22, n. 3, p. 179-191, set./dez. 2012.

STRASSER, Carly et al. Primer on data management: what you always wanted to know: but were afraid to ask. **DataONE**, Oakland, CA, p. 1-11, fev. 2012.

HUMANIDADES DIGITAIS:
visualização da produção científica

DIGITAL HUMANITIES:
visualization of scientific production

Mirelys Puerta Díaz[1], Edgar Bisset Alvarez[2], Silvana Aparecida Borsetti Gregorio Vidotti[3]
(1) Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP, Campus Marília), mireblue18@gmail.com, mirelys.puerta@unesp.br
(2) Universidade Federal de Santa Catarina, Florianópolis, edgar.bisset@ufsc.br
(3) Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP, Campus Marília), silvana.vidotti@unesp.br

Resumo:
O presente artigo aborda as Humanidades Digitais através da produção científica publicada nas bases de dados científicas nacionais e internacionais. No estudo, utilizou-se o método da bibliometria aplicada para identificar a produção científica indexada sobre o objeto pesquisado. A pesquisa é descritiva e quali-quantitativa. Utilizou-se como fonte, a Base de dados de Periódicos em Ciência da Informação e documentos recuperados com o Google Acadêmico através do software Publish or Perish, visando verificar a produção teórica das HD e as métricas de interesse. Aplicou-se para a visualização das métricas, o aplicativo *RAW Graphs*, baseado na web, trata-se de um framework de código aberto utilizado para a visualização de dados complexos. Com isso, foi possível constatar que no Brasil, os estudos sobre humanidades digitais na Ciência da Informação foram publicados somente a partir do ano 2014, enquanto no contexto internacional, as publicações já vinham sendo publicadas desde aproximadamente o ano 2000.
Palavras-chave:
Humanidades digitais. Visualização de dados. Análise métrica da produção científica.

Abstract:
This article addresses the Digital Humanities through the scientific production published in national and international scientific databases. In the study, the applied bibliometric method was used to identify the indexed scientific production on the researched object. The research is descriptive and qualitative-quantitative. It was used as source, the Database of Periodicals in Information Science and documents retrieved with Google Academic through the software Publish or Perish, to verify the theoretical production of the HD and the metrics of interest. It was applied to the visualization of the metrics, the web-based RAW Graphs application, it is an open-source framework used for the visualization of complex data. With this, it was possible to verify that in Brazil, studies on digital humanities in Information Science were only published as from 2014, while in the international context, publications had been published since about 2000.
Keywords:
Digital Humanities. Data visualization. Scientific production metric analysis.

1 INTRODUÇÃO

Nos últimos anos tem se observado um crescente avanço das pesquisas que envolvem as Humanidades Digitais (HD), o que possibilitou definí-las como campo disciplinar, perfil formativo ou como saber ligado ao movimento do livre acesso. E, esta produção científica também se ampliou a nível internacional, e, de tal maneira, que na opinião de Rodríguez-Yunta (2012), atualmente as HD podem ser consideradas como uma *topic tag* de significado ambíguo, cujo exponencial crescimento deve motivar a reflexão dos profissionais da Ciências da Informação.

Deste modo, esta área de estudo tem se convertido em verdadeira fonte para o desenvolvimento de projetos apontados como ponto de encontro entre as Ciências da Computação, o Ambiente digital e as Ciências Humanas e Sociais, neste sentido, Ortega (2013, p.18):

> "[...] as Humanidades Digitais podem ser definidas como o espaço de convergência entre as ciências da computação, meio digital e as disciplinas humanísticas na busca de novos modelos interpretativos e novos paradigmas de conhecimento em concordância com as transformações operadas no âmbito da sociedade digital."

Entretanto, apesar da crescente popularidade deste campo, não existe consenso no que diz respeito às suas bases teórico-epistemológicas, muito menos, quanto à construção de uma definição que resuma a essência do que são as HD, como aliás, fica evidenciado, no levantamento feito por Jason Heppler[1] em 2015, que reuniu cerca de 800 definições diferentes sobre o que são as HDs.

2 OBJETIVOS

O presente trabalho tem como objetivo geral apresentar o comportamento da produção científica das Humanidades Digitais[2], a partir da análise dos dados coletados nas bases de dados Google Acadêmico e a Base de dados de Periódicos em Ciência da Informação (BRAPCI).

3 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa descritiva combina as abordagens qualitativa e quantitativa e aplica-se a análise métrica da produção científica para alcançar seu objetivo.

A coleta de dados foi realizada sem limitações na data de publicação na BRAPCI, e através do buscador Google Acadêmico utilizando o software *Publish or Perish* (PoP), onde limitou-se a busca aos anos 2000-2017. O uso de PoP se justifica porque a interface nativa de Google Acadêmico não é muito adequada para análises bibliométricas.

Por meio da coleta na BRAPCI, se pretende identificar, além da presença das Humanidades Digitais nas pesquisas da Ciência da Informação brasileira, o foco de suas abordagens. Com esse fim a busca foi realizada em língua portuguesa ("humanidades digitais"). O Método de pesquisa aplicado para a análise dos dados é o de análise bibliométrica da produção científica.

1 https://github.com/hepplerj/whatisdigitalhumanities/blob/master/dayofquotes_full.csv
2 Digital humanities, em sua versão em inglês.

A coleta, processamento e análise de dados em Google Académico, foi apoiada pelo uso da ferramenta PoP, que recupera e faz análise de citações acadêmicas obtidas de fontes como Google Acadêmico. Da variedade de fontes de dados que usa PoP (por exemplo, *Crossref*, Google Acadêmico, *Web of Science y Scopus*) foi escolhida a do Google Acadêmico (GA) porque é considerada uma útil ferramenta para pesquisa de literatura científica de forma ampla, em muitas disciplinas e fontes acadêmicas (artigos revisados por pares, teses, livros, resumos e artigos de editoras acadêmicas, sociedades profissionais, repositórios de pré-impressão, repositórios de universidades e outras organizações acadêmicas). O PoP facilita o cálculo de várias métricas[3] (HARZING, 2011), das quais, foram selecionadas para análise, as especificadas no Quadro 01:

Quadro 01

Métrica	Descrição
Anos de publicação	Primeiro e mais recente ano de publicação encontrado nos resultados atualmente selecionados.
Anos de citação	Número de anos, desde o primeiro ano encontrado nos resultados atualmente selecionados até o ano da consulta (geralmente o ano corrente).
Documentos	Número total de resultados atualmente selecionados.
Quantidade de citações recebidas	A soma das contagens de citações em todos os resultados atualmente selecionados.
Promedio de citações por año (citações/ano)	Número médio de citações por ano (ou seja, citações / anos de citação)
Promedio de citações por documento (Citações / documento)	A soma das contagens de citações em todos os artigos, dividida pelo número total de artigos. A média e o modo também são calculados e disponibilizados separadamente.
Citações / autor	*Número médio*[4] de citações por autor
Artigos / autor	*Número médio*[5] de artigos por autor.
Autores /documento	*Número médio de autores*[6] por artigo.
h-index[7]	Índice H de Hirsch
índice g[8]	Índice g de Egghe

Fonte: Research in International Management, 2017.

Os resultados obtidos com o programa ficam disponíveis na tela e também podem ser copiados para a área de transferência do Windows, para colar em outros aplicativos ou salvar em vários formatos de saída (*vector graphic (svg), image (png), data model (json)*) para futura referência ou análise posterior. Sendo que, a capacidade de exportar o código embutido em SVG e o modelo de

3 https://harzing.com/resources/publish-or-perish/manual/using/query-results/metrics

4 Calculado da seguinte forma: Para cada trabalho, sua contagem de citações é dividida pelo número de autores para esse artigo, para devolver a contagem de citações normalizada por autor para o artigo. As contagens de citações normalizadas são então somadas em todos os artigos para fornecer o número de citações por autor sobre o conjunto de resultados.

5 Calculado da seguinte forma: Para cada trabalho, a contagem de 1 / autor é calculada para fornecer a contagem de autor normalizada para o trabalho. A contagem normalizada dos autores é então resumida em todos os artigos para fornecer o número de artigos por autor.

6 Calculado como a soma do autor conta em todos os artigos, dividido pelo número total de artigos. A média e o modo também são calculados e disponibilizados separadamente.

7 Métrica que mede o impacto cumulativo da publicação de um pesquisador, observando a quantidade de citações que seu trabalho recebeu. O índice h foi proposto por J.E. Hirsch no ano 2005.

8 O índice g é um índice para quantificar a produtividade na ciência, com base no registro de publicação (uma métrica em nível de autor). O índice é calculado com base na distribuição de citações recebidas pelas publicações de um determinado pesquisador, de tal forma que dado um conjunto de artigos classificados em ordem decrescente do número de citações que eles receberam, o índice g é o maior número único de tal forma que o topo g artigos recebidos juntos pelo menos g 2 citações.

dados JSON é compatível com essa abordagem aberta, o que permite explorar novos usos para esses produtos, facilitando o surgimento de novos métodos de visualização dos dados inseridos.

Aplicou-se para a visualização dos dados resultantes da coleta com o PoP, o aplicativo baseado na web, *RAWGraphs*[9]; um framework de código aberto utilizado para a visualização de dados complexos, desenvolvido no Laboratório de Pesquisa *DensityDesign* (Politecnico di Milano), cria visualizações baseadas em vetores personalizados da biblioteca *d3.js*. O *RAWGraphs* funciona com valores separados por delimitador (ou seja, arquivos csv e tsv), bem como, com textos copiados e colados de outros aplicativos (por exemplo, *Microsoft Excel*, *TextWrangler*, *TextEdit*). E, com base no formato svg, as visualizações podem ser facilmente importadas e editadas com aplicativos de gráficos vetoriais (ou seja, *Adobe Illustrator, Inkscape*, entre outros) para refinamentos adicionais ou diretamente incorporadas em páginas da web. Sendo que, o aplicativo *Inskcape* foi utilizado para melhorar a aparência dos gráficos gerados pelo aplicativo web *RAWGraphs*.

RAWGraphs surge como uma das ferramentas mais favoráveis para usar tanto em termos de características operacionais quanto características funcionais para a visualização de dados (ATWOOD e REZNIK-ZELLEN, 2018, p18).

Neste estudo foram utilizadas as seguintes visualizações do *RAWGraphs:*

- Gráfico de Dispersão (*Scatter Plot*): facilita criar Círculos aninhados para representar hierarquias e comparar valores. Essa visualização é particularmente eficaz para mostrar a proporção entre elementos através de suas áreas e posição dentro de uma estrutura hierárquica.
- Círculo *Packing*: Círculos aninhados permitem representar hierarquias e comparar valores. Essa visualização é particularmente eficaz para mostrar a proporção entre elementos através de suas áreas e sua posição dentro de uma estrutura hierárquica.

Foram identificadas como limitações do presente estudo, erros ou omissões referentes ao uso do PoP no Google Acadêmico, as quais são mais frequentes do que no Scopus e no *Web of Science*. O uso do programa requer a limpeza da lista de publicações obtidas na busca, eis que, também coletou registros que não correspondem com a estratégia de busca. Além disso o Google Scholar limita seus resultados a 1000 registros, os quais são classificados por número de citações recebidas, portanto, os 1000 coletados pelo PoP são os resultados mais citados, sendo assim omitidos muitas publicações que ainda não foram citadas.

4 RESULTADOS

Verificou-se com o levantamento bibliográfico realizado na base de dados BRAPCI, que trouxe como resposta, apenas, um total de 9 documentos, dos quais, 5 foram publicados no ano de 2017, o raso aprofundamento da temática pela comunidade acadêmica no Brasil.

Sendo que, o principal foco das pesquisas desenvolvidas no cenário brasileiro abordam as humanidades digitais relacionadas à temáticas como, as tecnologias da informação e as comunicações, suas abordagens desde a Ciência da Informação, sua relação com a mediação, a web, a ciência aberta, entre outros que estão representados na nuvem de palavras da Figura 1:

9 O projeto, liderado e mantido pelo Laboratório de Pesquisas *DensityDesign* (Politecnico di Milano) foi lançado publicamente em 2013 e é considerado como uma das ferramentas mais importantes no campo da visualização de dados.

Figura 1: Visualização de nuvem de palavras-chave que aparecem nos documentos recuperados da Brapci,

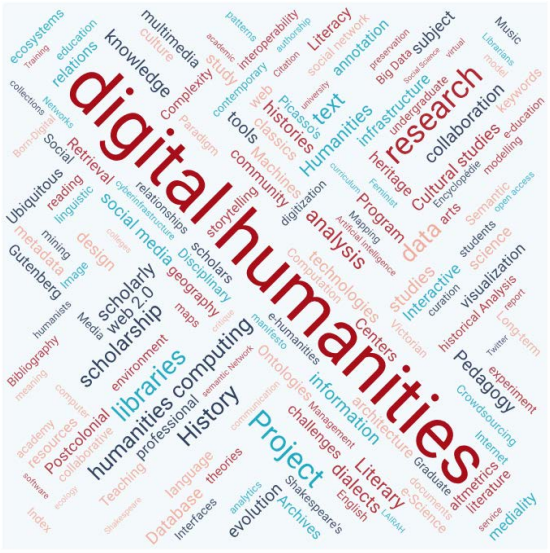
Fonte: Elaboração própria por meio da ferramenta on-line wordart.com.

Já a busca realizada no Google Acadêmico através do software PoP, recuperou, um total de 883 documentos, dos quais 510 são livros e 373 artigos de periódicos.

Os trabalhos levantados para esta pesquisa têm as humanidades digitais e a sua relação com áreas como, as bibliotecas como principal foco, além disso, também as *humanities computing*, suas abordagens desde a História como ciência, sua relação com o fenômeno do *big data*.

Identifica-se na própria infraestrutura das pesquisas sobre as humanidades digitais, um interesse pela análise do seu próprio ecossistema, suas abordagens vinculadas a web 2.0, Internet, as redes sociais, assim como as diversas ferramentas para sua visualização. Resulta interessante a quantidade de estudos de caso que descrevem os projetos em humanidades digitais que surgiram nos últimos 17 anos; ficando demonstrado e representados na nuvem de palavras da Figura 2:

Figura 2: Visualização de nuvem de palavras que aparecem nos títulos dos documentos recuperados do PoP.



Fonte: Elaboração própria por meio da ferramenta on-line wordart.com.

E, a partir dos documentos recuperados pelo PoP foi possível realizar a fase de normalização dos dados devido à necessária limpeza da lista de publicações obtidas na busca. Porém, os resultados das métricas foram os especificados no Quadro 02:

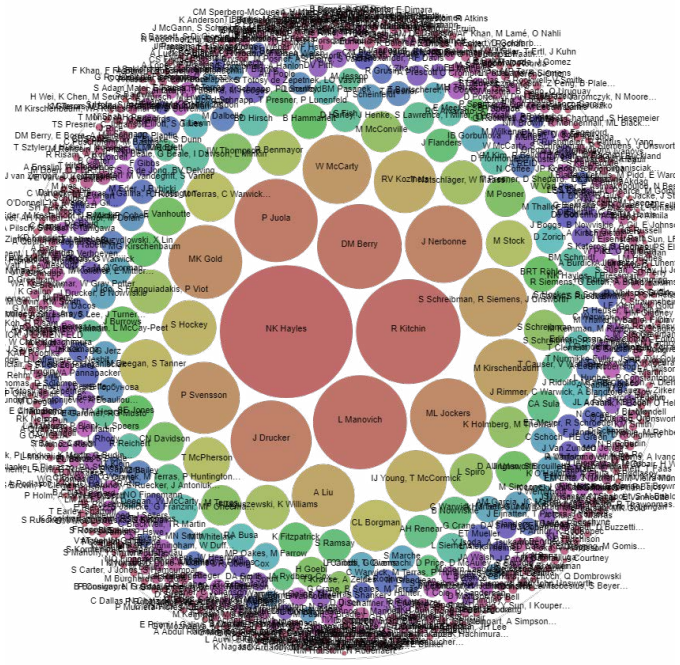
Quadro 02

Métrica	**Valor**
Anos de publicação	2000-2017
Anos de citação	17
Total de documentos	883
Quantidade de citações recebidas	34454
Promedio de citações por año (citações/ano)	453.34
Promedio de citações por documento (Citações / documento)	39.02
Citações / autor	29453.97
Artigos / autor	663.97
Autores / documento	1.80
h-index	74
índice g	164

Inicialmente foram recuperados um total de 1000 registros, sendo considerados para a análise 883, foram eliminados aqueles registros duplicados e que não respondiam à estratégia de busca definida, como se mostra no Quadro 2.

O número promedio das citações por ano, é uma métrica muito útil que possibilita a avaliação do impacto geral dos autores que publicam sobre esta temática em específico. As métricas individuais de Citações / autor recuperadas do PoP, apresenta o impacto de cada pesquisador, o que pode ser usado como uma alternativa aos *h-index* individuais. Um dado interessante que se percebe durante a análise, foi que os livros estão dentro as publicações com maior promedio de citação pelos autores.

Figura 3: Visualização da métrica de autores e total de citações recebidas.

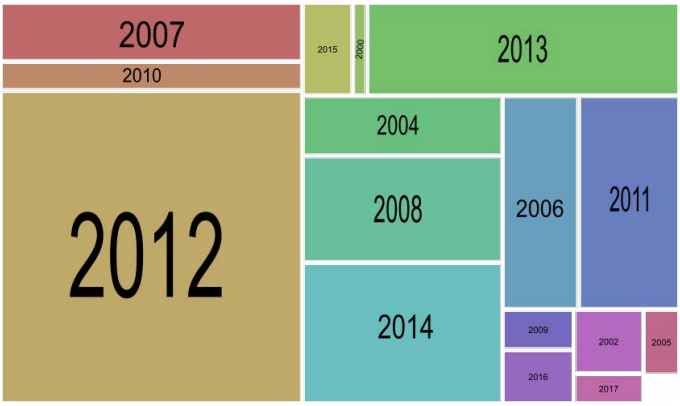


Fonte: Elaboração própria por meio da ferramenta RAWGraphs.

A Figura 3 demonstra que os autores que mais citações receberam (NK Hayles, NK Hayles, L Manovich,R Kitchin,P Juola, entre outros) encontram-se no centro da visualização, o que significa que quanto maior o número de citações, maior raio de cobertura dos autores que estão localizados no centro do gráfico.

O ano de maior quantidade de publicações sobre humanidades digitais foi o 2012 com um total de 166 trabalhos; seguido do ano 2013 com 122 publicações e 2014 com 103. Um fato interessante que nos últimos 3 anos diminuiu consideravelmente a produção científica de maior impacto sobre a temática segundo a fonte consultada em relação a anos anteriores. Do ano 2015 ao 2017 diminuiu a metade do total de publicações.

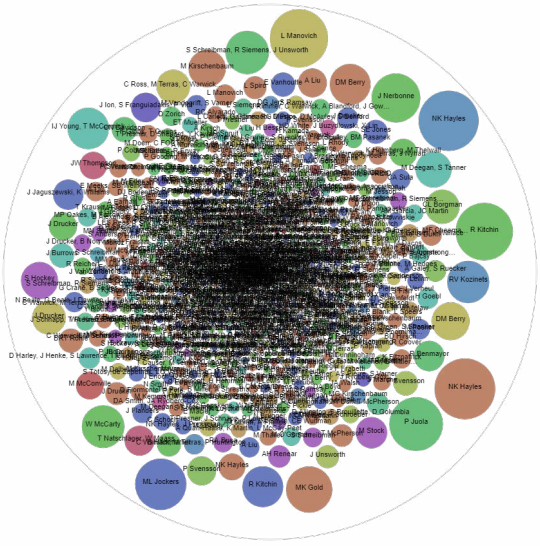
Figura 4: Visualização da métrica dos totales de produções científicas por ano.



Fonte: Elaboração própria por meio da ferramenta RAWGraphs.

A Figura 4 mostra a relação das dimensões ano de publicação e o total de citas recibidas por ano. Quanto maior o número de citações maior raio de cobertura dos anos que estão localizados no gráfico. Essa visualização das métricas mencionadas denota, com excepção do ano 2011 e 2008, que os três anos mais produtivos foram os mais citados.

Figura 5: Visualização das métricas do total de citações recebidas pelos autores e o GSR (*Google Scholar Rank*)



Fonte: Elaboração própria por meio da ferramenta RAWGraphs.

A visualização da Figura 5 denota que, os 5 trabalhos mais citados não aparecem nas 10 primeiras posições do ranking do Google Acadêmico, dos quais o 50% deles são publicações do ano 2012. Uma relação detalhada dos dados em relação aos autores mais citados está listada abaixo:

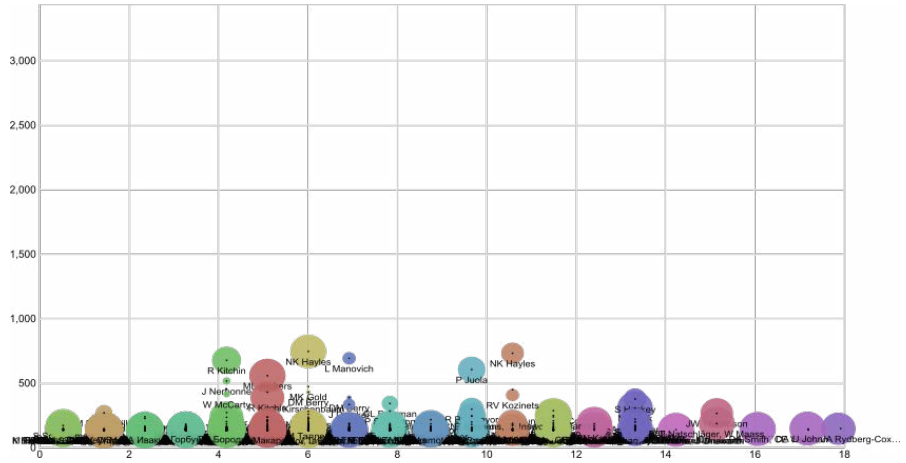
- NK Hayles (citações= 762; ano= 2012; GSRank =848)
- NK Hayles (citações= 746; ano= 2007; GSRank =315)
- L Manovich (citações= 700; ano= 2011; GSRank =95
- R Kitchin (citações= 682; ano= 2014; GSRank =542)
- P Juola (citações= 598; ano= 2008; GSRank =485)
- ML Jockers(citações= 543; ano= 2013; GSRank =861)

Identificou-se que somente o 25 % dos 10 primeiros resultados do Google Acadêmico estão em acesso aberto e disponíveis para a comunidade científica.

Essa visualização das métricas mencionadas, denota, que não necessariamente os artigos mais citados pelos autores (exemplo. Cuddon, Hayles e Röhle) são colocados nos primeiros 10 resultados da pesquisa feita pelos usuários no Google Acadêmico. Outro fato interessante é que, os autores com mais citações são os do ano 2012, como foi identificado na Figura 3.

Também se verifica, que as publicações correspondentes a 2012 estão entre as mais citadas.

Figura 6 Visualização de métrica da idade das publicações e do total de citações recebidas.



Fonte: Elaboração própria por meio da ferramenta RAWGraphs.

O gráfico da Figura 6, mostra a correlação entre a dimensão da idade das publicações e o total de citações recebidas. O Eixo X= idade das publicações (0-18); Eixo Y= total de citações recebidas por autor (0-762); Tamanho radio= GSRank (0-980); cor= autores).

No gráfico da Figura 6 a posição dos círculos que estão mais perto do Eixo Y são aqueles trabalhos que possuem menor tempo de publicação. Além disso, é preciso destacar que aquelas publicações que estão posicionadas mais perto do eixo Y, e que apresentam maior número de citações, coincidem com os trabalhos melhor posicionados no GSRank.

A Figura 6 mostra também o fato que das 7 pesquisas mais citadas e com mais de 500 citações recebidas (NK Hayles do 2011 e do ano 2007, L Manovich, R Kitchin, ML Jockers, P Juola e J Nerbonne) o 57% são das mais actuais, pois têm somente 6 anos de publicados, como é o caso das publicações de NK Hayles do 2011, R Kitchin, J Nerbonne e ML Jockers.

5 CONSIDERAÇÕES FINAIS

Foi possível concluir que o uso de softwares de visualização de dados nas análises métricas é comum na área da Ciência da Informação, embora, as configurações dos modelos complexos e altamente personalizados ainda sejam difíceis para os profissionais que não possuem as competências necessárias durante o desenvolvimento da pesquisa, o que demandaria uma quantidade significativa de tempo para correto domínio das ferramentas.

Também que, as principais publicações recuperadas sobre o tema estudado, não são oriundas de fontes ou tipologias documentais tradicionais de análise métrica da produção científica, e que os livros, são as principais fontes do foco de estudo.

Comportamentos interessantes foram identificados em relação às métricas de citações recebidas de artigos científicos de fontes alternativas às de sua publicação original, como é o caso das redes sociais acadêmicas (ex. academia.edu**,** ResearchGate). Este fato demonstra a tendência dos pesquisadores em compartilhar suas publicações e ganhar maior visibilidade para seu trabalho nestas plataformas, para além de sua publicação em periódicos científicos.

O ano de 2012 foi identificado como o ano mais produtivo em humanidades digitais e cujas publicações foram as mais citadas.

Identificou-se que a tendência geral é de que os documentos mais citados não apareçam nas 100 primeiras posições do ranking acadêmico do Google.

REFERÊNCIAS

ADAMS, J.L.; GUNN, K.B. Digital Humanities: Where to Start. **College & Research Libraries News**, v. 73, n. 9, p. 536-539, 569, 10 2012.

ATWOOD,T. P., e REZNIK-ZELLEN, R. Using the Visualization Software Evaluation Rubric to explore six freely available visualization applications. **Journal of eScience Librarianship**, v. 7, n.1, e1122 p. 1-20, 2018. Recuperado de: <https://escholarship.umassmed.edu/jeslib/vol7/iss1/3/>. Acessado em: 15 novembro 2018

HARZING, Anne-Wil, et al. **The Publish or perish book**. Melbourne: Tarma software research, 2011.

MAURI, M., Elli, T., CAVIGLIA, G., UBOLDI, G., & AZZI, M. RAWGraphs: A Visualisation Platform to Create Open Outputs. In **Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter** (p. 28:1–28:5), 2017. New York, NY, USA: ACM. Recuperado de: <https://doi.org/10.1145/3125571.3125585>. Acessado em: 24 septembro 2018.

POOLE, A.H. The conceptual ecology of digital humanities. **Journal of Documentation**, Bradford, v. 73, n. 1, p. 91-122, 2017.

RAWGraphs Team. **Introduction to RAWGraphs**. Licensed under CC BY-NC-SA 4.0. Recuperado de: <https://rawgraphs.io/learning/introduction-to-rawgraphs/>. Acessado em: 23 septembro 2018.

VERBEKE, D. Digital Humanities and Libraries - DH 2014. **META: tijdschrift voor bibliotheek & archief**, v. 90, n. 7, p. 41-42, 2014.

WONG, S.H.R. Digital Humanities: What Can Libraries Offer? **Portal : Libraries and the Academy**, Baltimore, v. 16, n. 4, p. 669-690, 10 2016.

ORTEGA, N. R. Humanidades Digitales, Digital Art History y cultura artística: relaciones y desconexiones. **ARTNODES: Revista de Arte, Ciencia y Tecnología**, Catalunya, n. 13, p. 16-25, 2013.

RODRÍGUEZ-YUNTA, L. Humanidades digitales, ¿una mera etiqueta o un campo por el que deben apostar las ciencias de la documentación?. **Anuário ThinkEPI**, v. 7, pp. 37-43, 2013.

IDENTIFICAÇÃO DE ENTIDADES DESTAQUE PARA A MELHORIA DA VISUALIZAÇÃO NA ANÁLISE DE GRAFOS

FEATURED ENTITY IDENTIFICATION TO IMPROVE THE GRAPH ANALYSIS VISUALIZATION

Roberto Zaina[1], Vinícius Faria Ramos Culmant[1], Gustavo Medeiros de Araujo[1]
(1) Universidade Federal de Santa Catarina, R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900, rzaina@gmail.com, v.ramos@ufsc.br, gustavo.araujo@ufsc.br

Resumo:
A proposta do presente estudo é a de desenvolver um método para a identificação automática de elementos relevantes em Relatórios de Inteligência Financeira. Então, inicialmente, explicamos a adoção de uma métrica específica para indicar suspeitas de lavagem de dinheiro, que é o uso de interpostas pessoas, comumente conhecidas como "laranjas". Uma forma de se identificar tais pessoas é verificar coincidências de pessoas como empregadas e sócias de empresas que operaram entre si. Detalhamos a metodologia adotada, que se iniciou com a tabulação das operações financeiras do Relatório de Inteligência Financeira e depois com a coleta dos dados relativos aos quadros societários das empresas e dos vínculos empregatícios dos sócios. Posteriormente, foi explicado o desenvolvimento de um *script* de extração, tratamento e carga dos dados em um programa de *Business Intelligence*, que trouxe como resultado uma tabela com as prováveis interpostas pessoas. Esta tabela, juntamente com os demais dados, foi importada para um programa de análise de vínculos, que permite visualizar os dados por meio de grafos. O resultado final foi um grafo que destaca as pessoas suspeitas de serem "laranjas" e, desta forma, facilita a análise do investigador.
Palavras-chave: Lavagem de Dinheiro. Interpostas Pessoas. Grafos.

Abstract:
The purpose of the present study is to develop a method for the automatic identification of relevant elements in Financial Intelligence Reports. So we initially explained the adoption of a specific metric to indicate suspicions of money laundering, which is the use of interposed people, commonly known as "oranges." One way to identify such people is to verify coincidences of people as employees and members of companies that operated among themselves. We detail the methodology adopted, which began with the tabulation of the financial operations of the Financial Intelligence Report and then with the collection of the data related to the companies' corporate staff and the employment relationships of the members. Subsequently, we explained the development of a script extracting, processing and loading the data into a Business Intelligence program, which resulted in a table with the likely interposed people. This table, along with the other data, was imported into a link analysis program, which allows visualizing the data through graphs. The final result was a graph highlighting the people suspected of being "oranges" and, thus, facilitates the researcher's analysis.
Keywords: Money laundry. Interposed People. Graphs.

1 INTRODUÇÃO

A investigação criminal é um conjunto de diligências que visam apurar a existência de um crime (BRAZ, 2013). Na investigação de lavagem de dinheiro o que se busca apurar é se bens ou valores foram ocultados ou dissimulados para esconder sua origem "suja".

Pela lavagem, o dinheiro proveniente de atividades ilícitas é introduzido nos circuitos financeiros legais, por meio de complexas operações que promovem a desvinculação da origem ilícita dos valores oriundos do crime (OLIVEIRA, 2012).

Uma das principais informações usadas nas investigações de lavagem é o Relatório de Inteligência Financeira (RIF), que é produzido pelo Conselho de Controle de Atividades Financeiras (COAF) e é resultado da análise de operações suspeitas de lavagem ou de financiamento do terrorismo (COAF, 2017).

No âmbito policial, as metodologias usadas para análise do RIF dependem totalmente da interpretação humana, mesmo quando se utilizam ferramentas tecnológicas.

Isto porque as ferramentas são usadas meramente para a visualização dos dados como, por exemplo, em planilhas ou diagramas, ficando o investigador encarregado de identificar o que é relevante no RIF, baseado em seu conhecimento e experiências profissionais.

2 OBJETIVOS

O objetivo geral deste estudo é desenvolver um método para a identificação automática de elementos relevantes em Relatórios de Inteligência Financeira.

E os objetivos específicos são:

- Estabelecer, com base em tipologias de lavagem de dinheiro, as métricas para a definição de elementos relevantes;
- Executar programa para processar as métricas nos dados tabulados;
- Executar programa para visualizar grafos sobre os dados processados.

3 PROCEDIMENTOS METODOLÓGICOS

O primeiro passo foi selecionar um critério para estabelecer relevância de elementos contidos nos Relatórios de Inteligência Financeira.

A escolha do critério foi norteada pelas tipologias descritas na publicação do artigo "Casos & Casos" do COAF, (COAF, 2016). (COAF, 2016) descreve 61 *modus operandi* de lavagem de dinheiro sobre recursos oriundos de crimes como corrupção, desvio de recursos públicos, contra o sistema financeiro, sonegação fiscal, tráfico de drogas e tráfico de pessoas.

Pela análise das tipologias foi selecionado um fenômeno que aparece em vários tipos de crimes, especialmente os de lavagem de dinheiro, que é a utilização de interpostas pessoas, também conhecidas como "laranjas",

As interpostas pessoas são as que têm sua documentação usada para, por exemplo, participar de empresas, abrir contas bancárias ou firmar contratos. Podem ser pessoas que não têm conhecimento de tal situação ou, então, sabem do uso de seus documentos, mas permitem isto em troca de algum tipo de vantagem financeira.

Um indicativo de interposta pessoa em uma empresa, é quando primeiro ela era empregada e depois se torna sócia na mesma empresa ou em outra empresa relacionada.

Este foi o critério selecionado para o presente estudo, ou seja, identificar pessoas que constam como empregadas e sócias de uma mesma empresa ou de empresas relacionadas. Neste caso, para estabelecer o que são "empresas relacionadas" decidimos usar aquelas que constam em operações financeiras de um mesmo RIF.

Estabelecidos os critérios para identificação de interpostas pessoas, o passo seguinte é a estruturação dos dados, a partir de um RIF pre-selecionado e com suas informações anonimizadas, por conta do sigilo policial (Apêndice A).

Considerando que o RIF é disponibilizado gerado em formato PDF, não estruturado, então foi necessário tabular suas informações em planilha. O resultado desta tabulação foi gravado como "OPERACOES_RIF.xlsx" (Apêndice B).

Os dados do RIF foram cruzados com outros dados advindos de sistemas corporativos do Governo Federal, por meio da ferramenta de BI *Qlikview*[1]. Criamos um novo documento no *Qlikview* denominado "RIF interpostas pessoas.qvw". Depois, os dados da planilha com as operações do RIF foram carregados pelo *script [OPERACOES RIF]* (Apêndice E).

Após esta primeira carga de dados, foi necessário avaliar quais os dados ainda seriam necessários para conseguir identificar as possíveis interpostas pessoas.

Em síntese, seriam necessários dados relativos aos quadros societários das empresas relacionadas no RIF e, em seguida, os vínculos empregatícios dos seus sócios.

Para tanto, decidimos testar alguns sistemas disponíveis na Polícia Federal, como os que contém dados registrados na Receita Federal, no Cadastro Nacional de Empresas, no Ministério do Trabalho e no Ministério da Previdência Social.

Como resultado destes testes, para encontrar os sócios das empresas foi escolhido um sistema da Receita Federal (RFB) que permite consultar pelo Cadastro Nacional de Pessoa Jurídica (CNPJ).

Para o histórico trabalhista dos sócios foi escolhido do Ministério da Previdência Social (MPS) que permite consultar pelo Cadastro de Pessoa Física (CPF).

Partindo para a consulta no sistema da RFB, surgiu uma dificuldade na seleção dos CNPJs que constam nas operações do RIF. Isto porque nas operações também podem constar CPFs. Dependendo da quantidade de operações do RIF, fazer manualmente esta seleção de CNPJs levaria muito tempo.

Então, a solução foi programar para que o próprio *Qlikview* fizesse a extração dos CNPJs, que são os valores com mais de 11 caracteres, que pode ser visto pelo *script [EXTRACAO CNPJ RIF]* (Apêndice E):

O resultado foi a criação de uma nova tabela "EXTRACAO CNPJ RIF", contendo somente os CNPJs que realizaram operações.

1 Ver mais em: https://www.qlik.com/pt-br

Depois, estes CNPJs foram consultados no sistema da Receita Federal (RFB), e o resultado foi estruturado em um arquivo denominado "CNPJ_RFB.csv", que está mostrado no Apêndice C. Posteriormente o arquivo foi carregado pelo *script [SOCIETARIO]* (Apêndice E).

O arquivo carregado conteve os dados das operações financeiras do RIF e, também, com os quadros societários das empresas que realizaram as operações.

Então, o próximo passo foi o de levantar os vínculos empregatícios dos sócios destas empresas.

Porém, como no campo de sócios também podem constar outras empresas, foi necessária a extração de sócios com CPFs, que são os valores com menos de 12 caracteres, que pode ser visto pelo *script [EXTRACAO CPF SOCIOS]* (Apêndice E):

O resultado da extração foi a criação de uma nova tabela "EXTRACAO CPF SOCIOS", contendo somente CPFs de sócios de empresas que realizaram operações.

Os CPFs extraídos pelo *script [EXTRACAO CPF SOCIOS]* foram consultados no sistema do Ministério da Previdência Social (MPS) e o resultado foi estruturado em um arquivo "CPF_CNIS.xls", que está mostrado no Apêndice D.

O arquivo resultado da consulta foi carregado no *Qlikview* com o seguinte *script [EMPREGATICIO]* (Apêndice E).

O script *[EMPREGATICIO]* retornou todos os dados necessários para a identificação das prováveis interpostas pessoas: operações financeiras do RIF, quadros societários das empresas e vínculos empregatícios dos sócios destas empresas.

Estes dados foram gravados no arquivo do *Qlikview* em 5 tabelas, conforme o modelo de dados mostrado na Figura 1:

Figura 1: Modelo de dados da 1ª carga

Considerando que se objetiva identificar pessoas que foram empregadas e sócias de empresas citadas no RIF, a próxima questão foi de verificar se existem registros de CNPJs na tabela "EMPREGATICIO" que também constam na tabela "SOCIETARIO".

Após o emprego de várias funções diferentes, escolhemos a *inner join*, que traz como resultado somente os registros existentes em ambas as tabelas. Com o novo *script [INTERPOSTA]*, o modelo de dados ficou conforme a Figura 2:

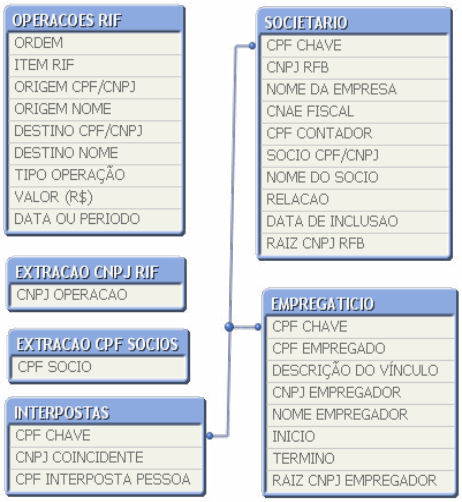
Figura 2: Modelo de dados da 2ª carga

O próximo passo foi o de estabelecer o relacionamento entre as tabelas, o que no *Qlikview* é feito pela padronização dos nomes dos campos a serem usados como "chaves".

Neste caso, adotou-se como chave o CPF das pessoas. Outra importante alteração foi considerar somente a raiz do CNPJ (8 primeiros algarismos), para unificar empresas do mesmo grupo econômico (matriz e filiais) (*script [SOCIETARIO],* Apêndice E).

Com estas alterações o modelo de dados ficou como mostra a Figura 3:

Figura 3: Modelo de dados da 3ª carga



Com este modelo de dados finalmente é possível a criação de objetos para consultar as prováveis interpostas pessoas.

4 RESULTADOS

O principal resultado foi o desenvolvimento de um método, por meio do programa *Qlikview*, que permite carregar e cruzar dados tabulados dos Relatórios de Inteligência Financeira e, ainda, de outros dados coletados, como os quadros societários das empresas que realizaram operações suspeitas e, por fim, os vínculos empregatícios dos sócios destas empresas.

Como resultado do processamento do *script* nos dados do "RIF modelo", chegou-se a Tabela 1, que mostra 04 CPFs suspeitos de serem de interpostas pessoas:

Tabela 1: CPFs dos possíveis "laranjas"

CPF INTERPOSTA PESSOA
311133387
2288877700
34545476598
52627277933

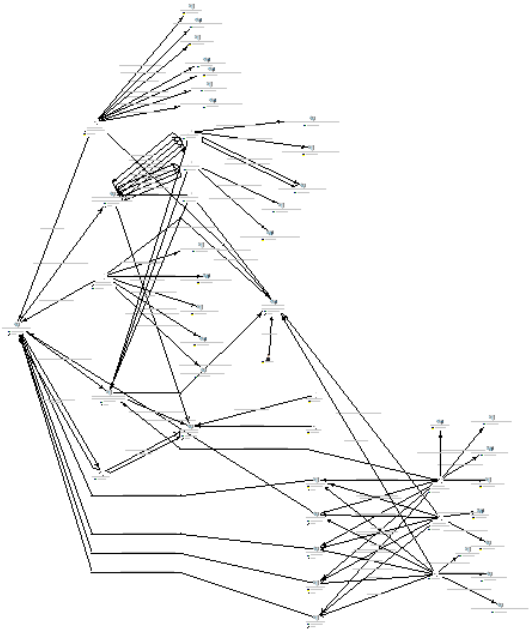
Este resultado pode ser aproveitado em outros programas usados pelos investigadores, especialmente os que permitem a "análise de vínculos", que mostram os dados por meio de grafos.

A análise de grafos é o "retrato gráfico de dados investigativos, feito para facilitar o entendimento de grandes volumes de dados, possibilitando aos investigadores entender as conexões entre indivíduos" (COADY, 1985 apud SPARROW, 1991, tradução nossa).

Para a criação dos grafos utilizamos o programa *IBM i2 Analyst's Notebook*[2], que permite importar dados tabulados. Foram importadas as planilhas usadas neste estudo e se definiu como um dos atributos das pessoas se ela é uma "interposta pessoa", conforme o resultado da Tabela 1.

O resultado da importação foi o grafo mostrado na Figura 4.

Figura 4: Grafo inicial

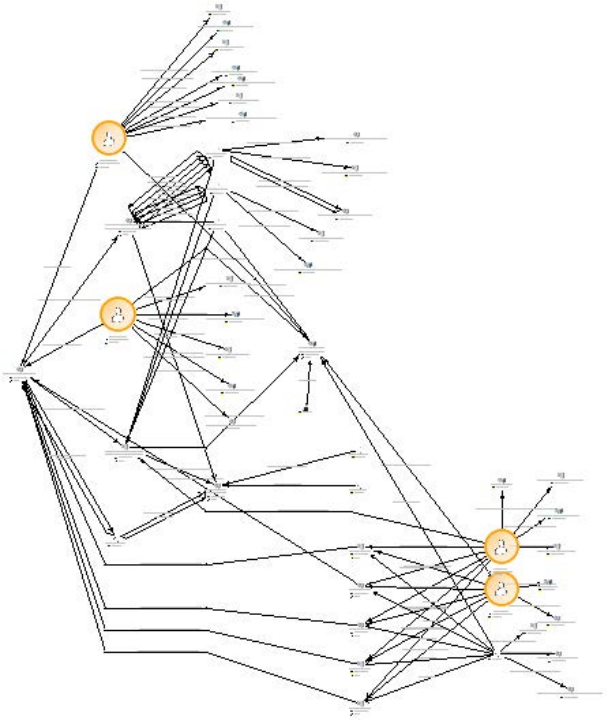


2 Ver mais em: www.ibm.com/br-pt/marketplace/analysts-notebook

Por enquanto, só pela análise visual da Figura 4 não se consegue perceber quais são as prováveis interpostas pessoas.

Contudo, considerando que foi definido um atributo "interposta pessoa", utiliza-se um recurso de formatação condicional para destacar tais ocorrências no grafo. O resultado disto é o grafo da Figura 5.

Figura 5: Grafo destacando os possíveis "laranjas"



Ao comparamos as Figura 4 e 5 fica evidente como o destaque das entidades do tipo "interposta pessoa" melhora a visualização do grafo e automatiza a detecção dos prováveis "laranjas" pelo investigador.

5 CONSIDERAÇÕES FINAIS

Com o objetivo de identificar elementos relevantes em Relatórios de Inteligência Financeira, escolhemos verificar a possibilidade de reconhecer automaticamente possíveis "laranjas" como sócias de empresas que realizam operações suspeitas.

Então, foi desenvolvido um método que envolve a coleta, a estruturação, o tratamento e a análise de dados relativos à RIF, quadros societários e vínculos trabalhistas.

Este estudo considerou uma tipologia bem específica, que é a utilização de prováveis "laranjas" para o cometimento de crimes e especificamente quando tais pessoas coincidem como empregadas e sócias de certas empresas.

Como principal resultado está o grafo mostrado na Figura 6, que ao destacar as entidades dos prováveis "laranjas" melhora significativamente a visualização do grafo e amplia instantaneamente a compreensão por parte do investigador que analisa o RIF.

A partir deste estudo pretendemos evoluir para estabelecer métricas de identificação de outras tipologias de lavagem de dinheiro e, assim, desenvolver novos métodos automáticos para análise de RIF.

REFERÊNCIAS

ANGELONI, Maria T.; REIS, Eduardo S. **Business Intelligence como Tecnologia de Suporte a Definição de estratégias para melhoria da qualidade do ensino**. Enanpad, 2006. Disponível em: <www.anpad.org.br/enanpad/2006/dwn/enanpad2006-adid-0815.pdf>. Acesso em 29/09/2018.

BRASIL. COAF – **Conselho de Controle de Atividades Financeiras. Relatório de Atividades**, 2017. Disponível em: <http://www.fazenda.gov.br/centrais-de-conteudos/publicacoes/relatorio-de-atividades/arquivos/relatorio-de-atividades-coaf-2017.pdf/view>. Acesso em: 29/09/2018.

BRASIL. LEI Nº 9.613, DE 3 DE MARÇO DE 1998. Disponível em: <http://www.planalto.gov.br/ccivil_03/LEIS/L9613.htm>. Acesso em: 29/09/2018.

BRASIL. Ministério da Fazenda. Conselho de Controle de Atividades Financeiras. Casos & Casos - Coletânea Completa de Casos Brasileiros de Lavagem de Dinheiro. Brasília: COAF, 2016. Disponível em: <https://www.fazenda.gov.br/centrais-de-conteudos/publicacoes/casos-casos/arquivos/casosecasos_coletanea-completa_setembro2016.pdf>. Acesso em: 29/09/2018.

BRAZ, José Alberto Campos. **Investigação criminal**. Leya, 2013.

FERRO JÚNIOR, Celso Moreira; DANTAS, George Felipe de Lima. **A descoberta e a análise de vínculos na complexidade da investigação criminal moderna**. 2006. Disponível em: <http://www.egov.ufsc.br/portal/sites/default/files/anexos/13124-13125-1-PB.pdf>. Acesso em: 03/06/2018.

OLIVEIRA, Tarsis Barreto. O bem jurídico-penal no crime de lavagem de dinheiro. **Revista Esmat**, [S.l.], v. 4, n. 4, p. 269-299, set. 2012. ISSN 2447-9896.

SPARROW, Malcolm K. **The application of network analysis to criminal intelligence**: An assessment of the prospects. Social networks, v. 13, n. 3, p. 251-274, 1991.

APÊNDICE A – "RIF MODELO" - RELATÓRIO DE INTELIGÊNCIA FINANCEIRA (RIF) COM DADOS ANONIMIZADOS

1. A empresa X1 Fomento Mercantil Ltda, sediada em Brasília/DF, atuante no ramo de administração de consórcios, com capital social de R$ 100.000,00, foi objeto de comunicações de operações financeiras de que trata a Lei 9.613/98. Desde 02/12/2013, figuram como sócios Fulano da Silva e Maria Beltrana, que até março de 2014, eram funcionários da empresa Contan Logística Ltda, com rendas mensais de R$ 1.836,00 e R$ 1.522,00, respectivamente.

1.1. Conforme o comunicante, a X1 Fomento Mercantil Ltda movimentou, entre 01/09/2014 e 15/10/2014, o total de R$ 35.052.870,00, na conta corrente nº 12345, agência nº 5234, do Banco Financeiro S.A., da cidade de Taguatinga/DF.

1.1.1. Os créditos somaram R$ 17.570.373,00, provenientes de 339 TEDs e transferências entre contas, que tiveram como principais remetentes identificados:

.R$ 4.411.827,00 - Auto Posto 1 Ltda
.R$ 4.300.049,00 - Auto Posto 5 Ltda
.R$ 3.655.000,00 - Auto Posto 2 Ltda
.R$ 3.124.624,00 - Auto Posto 4 Ltda
.R$ 1.963.150,00 - Auto Posto 3 Ltda

1.1.2. Os débitos no mesmo período totalizaram R$ 17.482.497,00, dos quais R$ 49.712,00, utilizados para pagamentos diversos, R$ 100.000,00 sacados em espécie por Fabiana Siqueira e R$ 17.329.785,00, destinados para quitação de 178 TEDs e transferências entre contas, dos quais foram identificados os seguintes principais beneficiários:

- 8.572.000,00 - Logs Distribuidora de Petróleo Ltda
- 8.331.317,00 - Logs Logística e Serviços Ltda

1.2. A movimentação suspeita baseou-se nos seguintes enquadramentos:

- *movimentação de recursos incompatível com o patrimônio, a atividade econômica ou a ocupação profissional e a capacidade financeira do cliente; Banco Central do Brasil - Carta-Circular nº 3.542, art. 1º*
- *recebimento de recursos com imediata compra de instrumentos para a realização de pagamentos ou de transferências a terceiros, sem justificativa; Banco Central do Brasil - Carta-Circular nº 3.542, art. 1º*
- *existência de contas que apresentem créditos e débitos com a utilização de instrumentos de transferência de recursos não característicos para a ocupação ou o ramo de atividade desenvolvida pelo cliente; Banco Central do Brasil - Carta-Circular nº 3.542, art. 1º*

2. Fabiana Siqueira foi relacionada em comunicação automática que registrou saque em espécie no dia 06/10/2014 no valor total de R$ 100.000,00, na conta corrente nº 12345, agência nº 5234, do Banco Financeiro S.A., da cidade de Taguatinga/DF, titulada pela X1 Fomento Mercantil Ltda, conforme detalhado no Anexo I.

2.1. Fabiana Siqueira foi relacionada em outra comunicação automática que registrou depósito em espécie no dia 05/11/2012 no valor total de R$ 130.000,00, na conta corrente nº 23344, agência nº 6789, do Banco Atual S.A., da cidade de Brasília/DF, de titularidade da empresa Zel Petróleo Ltda, conforme detalhado no Anexo I.

3. Em comunicação de operação suspeita, entre 01/09/2014 e 15/10/2014, titulada pela empresa Logs Logística e Serviços Ltda, na conta corrente nº 554433, agência nº 2278, do Banco Financeiro S.A., da cidade de Brasília/DF, no valor de R$ 16.821.019,00, sendo R$ 8.410.617,00 em créditos e R$ 8.410.402,00 em débitos. Os débitos foram realizados por transferências entre contas, dos quais R$ 4.734.930,00 foram remetidos para Zel Petróleo Ltda e R$ 496.614,00 para Contan Logística Ltda.

4. Em comunicação de operação suspeita, entre 01/09/2014 e 15/10/2014, titulada pela empresa Logs Distribuidora de Petróleo Ltda, na conta corrente nº 887766, agência nº 2278, do Banco Financeiro S.A., da cidade de Brasília/DF, no valor de R$ 16.821.019,00, sendo R$ 8.653.590,00 em créditos e R$ 8.653.301,00 em débitos. Os débitos foram realizados por transferências entre contas, dos quais R$ 4.637.200,00 foram remetidos para Zel Petróleo Ltda.

4.1. As empresas Logs Distribuidora de Petróleo Ltda e Logs Logística e Serviços Ltda pertencem a Cicrano Ferreira e Beltrano de Sousa. De acordo com informações descritas na mídia, datada em 14/07/2014, a Logs Distribuidora de Petróleo Ltda foi citada em investigação da Polícia Civil do Distrito Federal, por envolvimento em apreensão de solventes que visava a adulteração de combustíveis:

> *"Policiais da Delegacia de Crimes Fazendários prenderam, em flagrante, na madrugada desta sexta-feira, os motoristas de um caminhão carregado com 40 mil litros de solvente que, supostamente, seriam usados na adulteração de gasolina de uma distribuidora de combustíveis de Brasília. Os dois responderão a processo por estelionato. O produto ainda passará por análise.*
> *A carga apreendida na divisa entre Brasília e Taguatinga, segundo a Delegacia de Crimes Fazendários, vinha de Goiânia, e teria como destino a cidade de Sobradinho/DF. Segundo depoimento dos acusados, a carga seria entregue à empresa Logs Distribuidora de Petróleo Ltda., que tem sede em Brasília.*
> *Os motoristas também disseram à polícia que, semanalmente, traziam para a empresa 240 mil litros de solvente. Pelo volume, a polícia calcula que o lucro ilegal dessa operação seria de aproximadamente R$ 1 milhão por semana."*

5. Cicrano Ferreira foi relacionado em comunicações de que trata a Lei 9.613/98. Em especial, no período de 01/10/2014 e 15/10/2014, ele foi relacionado como titular em 4 comunicações automáticas que registraram saques em espécie no valor total de R$ 700.000,00, realizados na conta corrente nº 887766, agência nº 2278, do Banco Financeiro S.A., da cidade de Brasília/DF, titulada pela empresa Logs Distribuidora de Petróleo Ltda, conforme detalhado no Anexo I.

6. Beltrano de Sousa foi relacionado em 3 comunicações automáticas que registraram saques em espécie no valor total de R$ 400.000,00, no período de 06/10/2014 e 10/10/2014, realizados na conta corrente nº 887766, agência nº 2278, do Banco Financeiro S.A., da cidade de Brasília/DF, titulada pela empresa Logs Distribuidora de Petróleo Ltda, conforme detalhado no Anexo I.

1 - FABIANA SIQUEIRA

1.1.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
RICARDO GIL	087.654.321-00	Responsável
FABIANA SIQUEIRA	858.756.543-53	Sacador
X1 FOMENTO MERCANTIL LTDA	19.555.999/0001-01	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO FINANCEIRO	TAGUATINGA/DF	AGÊNCIA Nº 5234	12345	06/10/2014	100.000,00

Informações Adicionais:

Ocorrências:
Saque em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 - Circular 3.461/09, art. 12 - II

1.2.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
ZEL PETROLEO LTDA	02.333.777/0001-13	Responsável
FABIANA SIQUEIRA	858.756.543-53	Depositante
ZEL PETROLEO LTDA	02.333.777/0001-13	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO ATUAL	GOIANIA/GO	AGÊNCIA Nº 6789	23344	05/11/2012	130.000,00

Informações Adicionais:

Ocorrências:
Depósito em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 – Circular 3.461/09, art. 12 - II

2 - CICRANO FERREIRA

2.1.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
CICRANO FERREIRA	011.444.333-40	Responsável
CICRANO FERREIRA	011.444.333-40	Sacador
LOGS DISTRIBUIDORA DE PETRÓLEO LTDA	07.555.222/0001-40	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO FINANCEIRO	BRASÍLIA/DF	AGÊNCIA Nº 2278	887766	01/10/2014	200.000,00

Informações Adicionais:

Ocorrências:
Saque em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 – Circular 3.461/09, art. 12 - II

2.2.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
CICRANO FERREIRA	011.444.333-40	Responsável
CICRANO FERREIRA	011.444.333-40	Sacador
LOGS DISTRIBUIDORA DE PETRÓLEO LTDA	07.555.222/0001-40	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO FINANCEIRO	BRASÍLIA/DF	AGÊNCIA Nº 2278	887766	03/10/2014	100.000,00

Informações Adicionais:

Ocorrências:
Saque em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 - Circular 3.461/09, art. 12 - II

2.3.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
CICRANO FERREIRA	011.444.333-40	Responsável
CICRANO FERREIRA	011.444.333-40	Sacador
LOGS DISTRIBUIDORA DE PETRÓLEO LTDA	07.555.222/0001-40	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO FINANCEIRO	BRASÍLIA/DF	AGÊNCIA Nº 2278	887766	07/10/2014	250.000,00

Informações Adicionais:

Ocorrências:
Saque em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 - Circular 3.461/09, art. 12 - II

2.4.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
CICRANO FERREIRA	011.444.333-40	Responsável
CICRANO FERREIRA	011.444.333-40	Sacador
LOGS DISTRIBUIDORA DE PETRÓLEO LTDA	07.555.222/0001-40	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO FINANCEIRO	BRASÍLIA/DF	AGÊNCIA Nº 2278	887766	10/10/2014	150.000,00

Informações Adicionais:

Ocorrências:
Saque em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 - Circular 3.461/09, art. 12 - II

3 - BELTRANO DE SOUSA

3.1.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
BELTRANO DE SOUSA	099.222.555-87	Responsável
BELTRANO DE SOUSA	099.222.555-87	Sacador
LOGS DISTRIBUIDORA DE PETRÓLEO LTDA	07.555.222/0001-40	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO FINANCEIRO	BRASÍLIA/DF	AGÊNCIA Nº 2278	887766	06/10/2014	200.000,00

Informações Adicionais:

Ocorrências:
Saque em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 - Circular 3.461/09, art. 12 - II

3.2.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
BELTRANO DE SOUSA	099.222.555-87	Responsável
BELTRANO DE SOUSA	099.222.555-87	Sacador
LOGS DISTRIBUIDORA DE PETRÓLEO LTDA	07.555.222/0001-40	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO FINANCEIRO	BRASÍLIA/DF	AGÊNCIA Nº 2278	887766	09/10/2014	100.000,00

Informações Adicionais:

Ocorrências:
Saque em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 - Circular 3.461/09, art. 12 - II

3.3.

Relacionados	CPF/CNPJ	Tipo do Envolvimento
BELTRANO DE SOUSA	099.222.555-87	Responsável
BELTRANO DE SOUSA	099.222.555-87	Sacador
LOGS DISTRIBUIDORA DE PETRÓLEO LTDA	07.555.222/0001-40	Titular

Comunicante	Local	Agência - Código	Conta	Data	Valor em R$
BANCO FINANCEIRO	BRASÍLIA/DF	AGÊNCIA Nº 2278	887766	10/10/2014	100.000,00

Informações Adicionais:

Ocorrências:
Saque em espécie de valor igual ou superior a R$ 100.000,00 (cem mil reais). Banco Central do Brasil - Carta-Circular 3.098/03 - Circular 3.461/09, art. 12 - II

APÊNDICE B – TABULAÇÃO DAS OPERAÇÕES FINANCEIRAS CONTIDAS NO "RIF MODELO"

ORIGEM CPF/CNPJ	ORIGEM NOME	DESTINO CPF/CNPJ	DESTINO NOME	TIPO OPERAÇÃO	VALOR (R$)	DATA OU PERIODO
01.222.888/0001-70	Auto Posto 1 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda	Transferência	4.411.827,00	01/09/2014 a 15/10/2014
01.555.999/0001-30	Auto Posto 2 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda	Transferência	3.655.000,00	01/09/2014 a 15/10/2014
01.777.222/0001-52	Auto Posto 3 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda	Transferência	1.963.150,00	01/09/2014 a 15/10/2014
01.888.444/0001-88	Auto Posto 4 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda	Transferência	3.124.624,00	01/09/2014 a 15/10/2014
01.222.777/0001-01	Auto Posto 5 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda	Transferência	4.300.049,00	01/09/2014 a 15/10/2014
19.555.999/0001-01	X1 Fomento Mercantil Ltda	858.756.543-53	Fabiana Siqueira	Saque	100.000,00	06/10/2014
19.555.999/0001-01	X1 Fomento Mercantil Ltda	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	Transferência	8.572.000,00	01/09/2014 a 15/10/2014
19.555.999/0001-01	X1 Fomento Mercantil Ltda	06.999.555/0001-06	Logs Logística e Serviços Ltda	Transferência	8.331.317,00	01/09/2014 a 15/10/2014
858.756.543-53	Fabiana Siqueira	02.333.777/0001-13	Zel Petróleo Ltda	Depósito	130.000,00	05/11/2012
06.999.555/0001-06	Logs Logística e Serviços Ltda	02.333.777/0001-13	Zel Petróleo Ltda	Transferência	4.734.930,00	01/09/2014 a 15/10/2014
06.999.555/0001-06	Logs Logística e Serviços Ltda	15.151.151/0001-09	Contan Logística Ltda	Transferência	496.614,00	01/09/2014 a 15/10/2014
07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	02.333.777/0001-13	Zel Petróleo Ltda	Transferência	4.637.200,00	01/09/2014 a 15/10/2014
07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	011.444.333-40	Cicrano Ferreira	Saque	200.000,00	01/10/2014
07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	011.444.333-40	Cicrano Ferreira	Saque	100.000,00	03/10/2014
07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	011.444.333-40	Cicrano Ferreira	Saque	250.000,00	07/10/2014
07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	011.444.333-40	Cicrano Ferreira	Saque	150.000,00	10/10/2014
07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	099.222.555-87	Beltrano de Sousa	Saque	200.000,00	06/10/2014
07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	099.222.555-87	Beltrano de Sousa	Saque	100.000,00	09/10/2014
07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	099.222.555-87	Beltrano de Sousa	Saque	100.000,00	10/10/2014

APÊNDICE C – DADOS SOCIETÁRIOS DAS EMPRESAS RELACIONADAS NO "RIF MODELO"

CNPJ	NOME DA EMPRESA	CNAE FISCAL	CPF CONTADOR	SOCIO CPF/CNPJ	NOME DO SOCIO	RELACAO	DATA DE INCLUSAO
01.222.777/0001-01	AUTO POSTO 5 LTDA	Posto de combustíveis	087.654.321-00	345.454.765-98	TICIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.222.777/0001-01	AUTO POSTO 5 LTDA	Posto de combustíveis	087.654.321-00	526.272.779-33	MEVIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.222.888/0001-70	AUTO POSTO 1 LTDA	Posto de combustíveis	087.654.321-00	345.454.765-98	TICIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.222.888/0001-70	AUTO POSTO 1 LTDA	Posto de combustíveis	087.654.321-00	526.272.779-33	MEVIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.555.999/0001-30	AUTO POSTO 2 LTDA	Posto de combustíveis	087.654.321-00	345.454.765-98	TICIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.555.999/0001-30	AUTO POSTO 2 LTDA	Posto de combustíveis	087.654.321-00	526.272.779-33	MEVIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.777.222/0001-52	AUTO POSTO 3 LTDA	Posto de combustíveis	087.654.321-00	345.454.765-98	TICIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.777.222/0001-52	AUTO POSTO 3 LTDA	Posto de combustíveis	087.654.321-00	526.272.779-33	MEVIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.888.444/0001-88	AUTO POSTO 4 LTDA	Posto de combustíveis	087.654.321-00	345.454.765-98	TICIO MORAIS	Excluido: 20/12/2014	11/02/2014
01.888.444/0001-88	AUTO POSTO 4 LTDA	Posto de combustíveis	087.654.321-00	526.272.779-33	MEVIO MORAIS	Excluido: 20/12/2014	11/02/2014
02.333.777/0001-13	ZEL PETROLEO LTDA	Distribuidora de combustíveis	858.756.543-53	938.743.646-87	MILTON WILLIAM	Socio	11/04/2011
02.333.777/0001-13	ZEL PETROLEO LTDA	Distribuidora de combustíveis	858.756.543-53	987.636.545-59	DIEGO TORQUATO	Administrador	11/04/2011
06.999.555/0001-06	LOGS LOGISTICA E SERVICOS LTDA	Serviços de apoio administrativo e logístico	654.365.645-09	011.444.333-40	CICRANO FERREIRA	Socio	21/08/2010
06.999.555/0001-06	LOGS LOGISTICA E SERVICOS LTDA	Serviços de apoio administrativo e logístico	654.365.645-09	099.222.555-87	BELTRANO DE SOUSA	Administrador	21/08/2010
07.555.222/0001-40	LOGS DISTRIBUIDORA DE PETROLEO LTDA	Distribuidora de combustíveis	654.365.645-09	011.444.333-40	CICRANO FERREIRA	Socio	25/07/2010
07.555.222/0001-40	LOGS DISTRIBUIDORA DE PETROLEO LTDA	Distribuidora de combustíveis	654.365.645-09	099.222.555-87	BELTRANO DE SOUSA	Administrador	25/07/2010
15.151.151/0001-09	CONTAN LOGISTICA LTDA	Serviços de contabilidade	123.777.222-99	087.654.321-00	RICARDO GIL	Socio	25/04/2014
15.151.151/0001-09	CONTAN LOGISTICA LTDA	Serviços de contabilidade	123.777.222-99	654.365.645-09	ANTONIO FERREIRA	Administrador	25/04/2012
19.555.999/0001-01	X1 FOMENTO MERCANTIL LTDA	Fomento mercantil	858.756.543-53	003.111.333-87	MARIA BELTRANA	Administrador	02/12/2013
19.555.999/0001-01	X1 FOMENTO MERCANTIL LTDA	Fomento mercantil	858.756.543-53	022.888.777-00	FULANO DA SILVA	Socio	02/12/2013

APÊNDICE D – DADOS EMPREGATÍCIOS DOS SÓCIOS DAS EMPRESAS DO "RIF MODELO"

IDENTIFICADOR ORIGEM	DESCRIÇÃO DO VÍNCULO	IDENTIFICADOR DESTINO	NOME EMPREGADOR	INICIO	TERMINO
00311133387	EMPREGADO (01/01/2013 a 01/03/2014) (CNIS)	15151151000109	CONTAN LOGÍSTICA LTDA	01/01/2013	01/03/2014
00311133387	EMPREGADO (01/11/1992 a 15/10/1993) (CNIS)	78080151000186	BOM FILE SERVICOS DE ALIMENTOS LTDA	01/11/1992	15/10/1993
00311133387	EMPREGADO (07/11/1994 a 10/05/1995) (CNIS)	76754018000185	CORRETO SERVICOS TEMPORARIOS LTDA	07/11/1994	10/05/1995
00311133387	EMPREGADO (11/05/1995 a 01/08/1995) (CNIS)	76555057000187	MONZA INDUSTRIA METALURGICA LTDA	11/05/1995	01/08/1995
00311133387	EMPREGADO (01/10/1994 a 12/1994) (CNIS)	75568011000100	VITORIA COMERCIO DE COMBUSTIVEIS LTDA	01/10/1994	01/12/1994
00311133387	EMPREGADO (01/12/1995 a 04/1996) (CNIS)	00147507000110	EMPONTO SERVICOS DE ENTREGAS ESPECIAIS LTDA	01/12/1995	01/04/1996
01144433340	EMPREGADO (20/08/1996 a 01/11/1996) (CNIS)	78654860000165	VETOR ORGANIZACAO DE RECURSOS HUMANOS S.A.	20/08/1996	01/11/1996
01144433340	EMPREGADO (01/08/1995 a 31/12/1997) (CNIS)	78050554000177	DIRETO EMPRESA DE SEGURANCA LTDA	01/08/1995	31/12/1997
01144433340	EMPREGADO (31/10/1996 a 05/12/1996) (CNIS)	76487685000551	COMPANHIA DE CIMENTO WINGS	31/10/1996	05/12/1996
01144433340	EMPREGADO (01/01/1998 a 25/11/1999) (CNIS)	76487685000551	COMPANHIA DE CIMENTO WINGS	01/01/1998	25/11/1999
02288877700	EMPREGADO (01/04/2013 a 01/03/2014) (CNIS)	15151151000109	CONTAN LOGÍSTICA LTDA	01/04/2013	01/03/2014
02288877700	EMPREGADO (29/09/2010 a 07/03/2013) (CNIS)	81188545000515	PRONTO ENGENHARIA DE PROJETOS LTDA	29/09/2010	07/03/2013
02288877700	EMPREGADO (02/02/2004 a 11/04/2005) (CNIS)	80841450000175	ASSOC DE PROT A MATERNIDADE DE CURITIBA	02/02/2004	11/04/2005
02288877700	EMPREGADO (03/12/2007 a 09/11/2009) (CNIS)	07754165000156	CENTRO INTEGRADO DE FONOAUDIOLOGIA	03/12/2007	09/11/2009
02288877700	EMPREGADO (01/10/2007 a 10/2007) (CNIS)	05784650000170	CAMARA MUNICIPAL	01/10/2007	01/10/2007
02288877700	EMPREGADO (01/05/2003) (CNIS)	05588855000177	PEREIRA CONSTRUCOES LTDA	01/05/2003	
02288877700	EMPREGADO (16/03/2006 a 03/05/2006) (CNIS)	04585468000175	RH TOTAL TRABALHO TEMPORARIO LTDA	16/03/2006	03/05/2006
02288877700	EMPREGADO (16/04/2002 a 08/12/2003) (CNIS)	01678156000155	LIQUIFAZ COMERCIO DE COMBUSTIVEIS LTDA	16/04/2002	08/12/2003
08765432100	EMPREGADO (19/04/2006 a 03/03/2009) (CNIS)	86854584006546	CONCRETUDE CIMENTOS BRASIL S.A.	19/04/2006	03/03/2009
08765432100	EMPREGADO (14/12/2005 a 18/04/2006) (CNIS)	86815681000178	REGRA SERVICO TEMPORARIO E GESTAO DE PESSOAS LTDA	14/12/2005	18/04/2006
08765432100	EMPREGADO (01/06/2011 a 20/08/2011) (CNIS)	08555548000184	MADEIREIRA PENEDO LTDA ME	01/06/2011	20/08/2011
09922255587	EMPREGADO (04/09/2006 a 20/04/2007) (CNIS)	85508765015586	TCX SUPERMERCADOS DO BRASIL LTDA	04/09/2006	20/04/2007
09922255587	EMPREGADO (24/04/2007 a 22/07/2007) (CNIS)	85504055000154	INSERVICE ADMINISTRACAO E REPRESENTACOES LTDA	24/04/2007	22/07/2007
34545476598	EMPREGADO (24/05/2013 a 15/11/2013) (CNIS)	06999555000106	LOGS LOGÍSTICA E SERVIÇOS LTDA	24/05/2013	15/11/2013
34545476598	EMPREGADO (01/12/2010 a 15/10/2012) (CNIS)	75045581000115	VIACAO ASA LTDA	01/12/2010	15/10/2012
34545476598	EMPREGADO (23/07/2007 a 08/03/2008) (CNIS)	07460455000140	QUANTO SERVICOS E LOGISTICA LTDA	23/07/2007	08/03/2008
34545476598	EMPREGADO (02/05/2013 a 10/10/2017) (CNIS)	06558855000104	FERREIRA COMERCIO DE MADEIRAS	02/05/2013	10/10/2017

34545476598	EMPREGADO (01/12/2009 a 14/04/2010) (CNIS)	00855751000118	AUTO POSTO DA LUZ LTDA.	01/12/2009	14/04/2010
52627277933	EMPREGADO (14/02/2013 a 11/10/2013) (CNIS)	15151151000109	CONTAN LOGÍSTICA LTDA	14/02/2013	11/10/2013
52627277933	EMPREGADO (01/02/1988 a 23/07/1989) (CNIS)	76545115005057	BANCO FINABRA S.A	01/02/1988	23/07/1989
52627277933	EMPREGADO (18/04/1996 a 18/12/1996) (CNIS)	75185587000187	AUTO VIACAO RAPIDO LTDA	18/04/1996	18/12/1996

APÊNDICE E – SCRIPT FINAL DO DOCUMENTO "RIF_INTERPOSTAS_PESSOAS.QVW"

```
/* Script criado para carga dos seguintes dados:
- Operações do RIF;
- Quadros societários das empresas que realizaram operações;
- Vínculos empregatícios dos sócios das empresas que realizaram operações. */

// Configurações gerais do documento:

SET ThousandSep='.';
SET DecimalSep=',';
SET MoneyThousandSep='.';
SET MoneyDecimalSep=',';
SET MoneyFormat='R$ #.##0,00;-R$ #.##0,00';
SET TimeFormat='hh:mm:ss';
SET DateFormat='DD/MM/YYYY';
SET TimestampFormat='DD/MM/YYYY hh:mm:ss[.fff]';
SET FirstWeekDay=6;
SET BrokenWeeks=1;
SET ReferenceDay=0;
SET FirstMonthOfYear=1;
SET CollationLocale='pt-BR';
SET MonthNames='jan;fev;mar;abr;mai;jun;jul;ago;set;out;nov;dez';
SET LongMonthNames='janeiro;fevereiro;março;abril;maio;junho;julho;agosto;setembro;outubro;novembro;dezembro';
SET DayNames='seg;ter;qua;qui;sex;sáb;dom';
SET LongDayNames='segunda-feira;terça-feira;quarta-feira;quinta-feira;sexta-feira;sábado;domingo';

// Carga de dados das Operações do RIF:

[OPERACOES RIF]:
LOAD
ORDEM,
[ITEM RIF],
KeepChar([ORIGEM CPF/CNPJ],'0123456789') as [ORIGEM CPF/CNPJ],
[ORIGEM NOME],
KeepChar([DESTINO CPF/CNPJ],'0123456789') as [DESTINO CPF/CNPJ],
[DESTINO NOME],
[TIPO OPERAÇÃO],
[VALOR (R$)],
[DATA OU PERIODO]
FROM *OPERACOES_RIF*.xlsx (ooxml, embedded labels, table is Plan1);

// Extração dos CNPJs que realizaram operações no RIF:

[EXTRACAO CNPJ RIF]:
LOAD
IF(LEN([ORIGEM CPF/CNPJ])>11,[ORIGEM CPF/CNPJ]) as [CNPJ OPERACAO]
Resident [OPERACOES RIF];
LOAD
IF(LEN([DESTINO CPF/CNPJ])>11,[DESTINO CPF/CNPJ]) as [CNPJ OPERACAO]
Resident [OPERACOES RIF];

// Carga de dados dos Quadros societários das empresas que realizaram operações:

SOCIETARIO:
LOAD
KeepChar(CNPJ,'0123456789') as [CNPJ RFB],
[NOME DA EMPRESA],
[CNAE FISCAL],
KeepChar([CPF CONTADOR],'0123456789') as [CPF CONTADOR],
KeepChar([SOCIO CPF/CNPJ],'0123456789') as [SOCIO CPF/CNPJ],
[NOME DO SOCIO],
RELACAO,
```

```
[DATA DE INCLUSAO],
KeepChar([SOCIO CPF/CNPJ],'0123456789') as [CPF CHAVE],
Left(KeepChar(CNPJ,'0123456789'),8) as [RAIZ CNPJ RFB]
FROM *CNPJ*.csv (txt, codepage is 1252, embedded labels, delimiter is ';', msq, filters(
Remove(Row, RowCnd(CellValue, 1, StrCnd(null)))));

// Extração dos CPFs dos sócios das empresas que realizaram operações no RIF:

[EXTRACAO CPF SOCIOS]:
LOAD
IF(LEN([SOCIO CPF/CNPJ])<12, [SOCIO CPF/CNPJ]) as [CPF SOCIO]
Resident SOCIETARIO;

// Carga de dados dos vínculos empregatícios dos sócios das empresas que realizaram operações:

EMPREGATICIO:
LOAD
KeepChar([IDENTIFICADOR ORIGEM],'0123456789') AS [CPF EMPREGADO],
[DESCRIÇÃO DO VÍNCULO],
KeepChar([IDENTIFICADOR DESTINO],'0123456789') AS [CNPJ EMPREGADOR],
[NOME EMPREGADOR],
INICIO,
TERMINO,
KeepChar([IDENTIFICADOR ORIGEM],'0123456789') as [CPF CHAVE],
Left(Right(000&KeepChar([IDENTIFICADOR DESTINO],'0123456789'),14),8) as [RAIZ CNPJ EMPREGADOR]
FROM *CNIS*.xls (biff, embedded labels, table is Plan1$);

// Identificação de "Interpostas Pessoas": CPFs relacionados à CNPJs que constam nas tabelas "SOCIETARIO" e "EM-
PREGATICIO"

INTERPOSTAS:
LOAD [RAIZ CNPJ RFB] as [CNPJ COINCIDENTE]
Resident SOCIETARIO;
Inner Join
LOAD[RAIZ CNPJ EMPREGADOR] as [CNPJ COINCIDENTE],
[CPF EMPREGADO] as [CPF INTERPOSTA PESSOA],
[CPF EMPREGADO] as [CPF CHAVE]
Resident EMPREGATICIO;
```

INFORMAÇÃO E PRIVACIDADE NA LEI GERAL DE PROTEÇÃO DE DADOS

INFORMATION AND PRIVACY IN THE GENERAL LAW OF DATA PROTECTION

Rosilene Paiva Marinho de Sousa[1], Fernando Antônio de Vasconcelos[2], Marckson Roberto Ferreira de Sousa[3]
(1) Centro Universitário de João Pessoa (UNIPÊ), adv.rpmarinho@gmail.com
(2) Centro Universitário de João Pessoa (UNIPÊ), fer.mengo@uol.com.br
(3) Universidade Federal da Paraíba (UFPB), mrck.sousa@gmail.com

Resumo:
As Tecnologias da Informação e Comunicação têm possibilitado a criação de uma grande quantidade de dados que necessitam ser devidamente controlados para a garantia da proteção dos direitos fundamentais da pessoa humana. O presente trabalho tem como escopo analisar a relação entre informação e privacidade no âmbito da Lei Geral de Proteção de Dados. Reflete sobre os instrumentos legais de amparo a proteção de dados no âmbito do ordenamento pátrio. Busca discorrer sobre os contornos que envolvem a referida lei na proteção da privacidade e o acesso à informação considerando o papel do consentimento como requisito para tratamento dos dados pessoais. Como metodologia adotar-se-á pesquisa bibliográfica e exploratória. Como resultados, examina como a referida lei coloca o acesso à informação como instrumento que permite a privacidade se manter como ponto central para a proteção dos direitos fundamentais dos indivíduos, concluindo que o controle torna-se instrumento de garantia dos direitos de liberdade, intimidade e privacidade.
Palavras-chave: Informação. Privacidade. Lei Geral de Proteção de Dados Pessoais.

Abstract:
Information and communication technologies have made it possible to create a large amount of data that needs to be properly monitored to guarantee the protection of the fundamental rights of the human Person. The present work is scoped to analyse a relationship between information and privacy under the general law of data Protection. It reflects on the legal instruments of protection of parental ordination. It seeks to discuss the contours that involve the law in the protection of privacy and access to information considering the role of consent as a requirement for the treatment of personal data. As a methodology, bibliographical and exploratory research will be adopted. As a result, it examines how an law places access to information as an instrument that allows a privacy to remain as a central point for the protection of the fundamental rights of individuals, concluding that control becomes a guarantee instrument the rights of freedom, intimacy and privacy.
Keywords: Information. Privacy. General Data Protection Law.

1 INTRODUÇÃO

As transformações advindas com as Tecnologias da Informação e Comunicação impactaram a forma de armazenar, processar e transmitir informações mudando rapidamente a vida de bilhões de pessoas. A proteção da privacidade tem sido uma preocupação constante em face do volume de dados produzidos com a massificação da utilização dos computadores.

O aumento do volume de dados tem proporcionado uma assimetria informacional em que ocorre a redução do conhecimento do cidadão sobre o uso de seus dados e dos efeitos advindos desse uso.

A dicotomia entre acesso à informação e a privacidade tem sido tema de discussões quanto aos contornos e limites de sua proteção. Surge dessa forma, a necessidade de disciplinar como e quando disponibilizar informações pessoais a terceiros. Nesse sentido, o sigilo posto como uma barreira ao acesso à informação passa a ser substituído pelo controle de seu acesso, permitindo assim que a privacidade possa se manter como ponto central para a proteção dos direitos fundamentais dos indivíduos, chegando-se ao arremate de que o controle torna-se instrumento de garantia dos direitos de liberdade, intimidade e privacidade.

Dessa forma, a relevância deste trabalho surge ao levar em consideração que a proteção sobre dados pessoais possa contribuir no controle do armazenamento, processamento e transmissão de informações, resguardando assim direitos fundamentais.

2 OBJETIVOS

Este trabalho tem como objetivo analisar a relação entre informação e privacidade no âmbito da Lei Geral de Proteção de Dados. Para isso, busca ainda refletir sobre os instrumentos legais de amparo a proteção de dados no âmbito do ordenamento pátrio, examinando contornos que envolvem a referida lei na proteção da privacidade e o acesso à informação, considerando o papel do consentimento como requisito para tratamento dos dados pessoais.

3 PROCEDIMENTOS METODOLÓGICOS

Do ponto de vista metodológico, trata-se de uma pesquisa bibliográfica e exploratória. A pesquisa bibliográfica consiste na seleção de informações bibliográficas (livros, artigos científicos, documentos, dentre outros), que possam contribuir para explicar o objeto de investigação (HENRIQUES; MEDEIROS, 2017, p.107). Já a pesquisa exploratória, segundo Severino (2007), visa levantar informações sobre um determinado objeto, delineando o campo de trabalho, bem como as condições de manifestação desse objeto.

4 RESULTADOS

Nos resultados são expostos discussões sobre como se apresentam a proteção de dados no ordenamento pátrio, seguido da análise da proteção da privacidade e do acesso à informação, destacando a importância do consentimento para o tratamento dos dados pessoais.

4.1 A Proteção aos Dados Pessoais no Ordenamento Pátrio

No panorama do ordenamento jurídico brasileiro que antecedia a regulação da proteção de dados, constatava-se a sua proteção por meio das garantias constitucionais da liberdade, intimidade e dignidade da pessoa humana, estando a privacidade na esfera dessas garantias.

A Constituição Federal (CF) considera invioláveis a vida privada, e a intimidade em seu artigo 5º, inciso X, bem como a interceptação de comunicações telefônicas, telegráfica ou de dados, no mesmo artigo, inciso XII. Ainda em âmbito constitucional, destacam-se os direitos de acesso, retificação e complementação de informações, com previsão no artigo 5º, incisos XIV e XXXIII (BRASIL,1988).

No âmbito infraconstitucional, destacavam-se a proteção de dados em leis específicas não havendo até então, uma proteção geral exclusiva sobre o tema.

Dentre essa proteção, o Código de Defesa do Consumidor, em seu artigo 43 determina a proteção do consumidor em vários aspectos envolvendo às informações pessoais presentes em banco de dados e cadastros dos consumidores (BRASIL, 1990).

Ainda no âmbito da proteção do consumidor, a Lei do Cadastro Positivo de nº 12.414/11, estabeleceu regras a partir de princípios de proteção de dados pessoais para tratamento de dados financeiros para formação de histórico de crédito (BRASIL, 2011a).

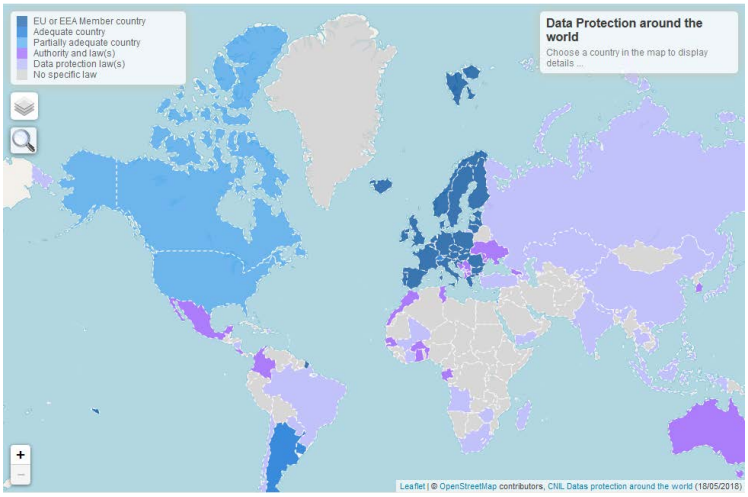
O Habeas Data foi regulado pela Lei nº 9.507/97, tendo como objetivo de garantir a retificação de dados pessoais em banco de dados de entidades governamentais ou de caráter público (BRASIL, 1997).

Outro instituto que pode ser citado, diz respeito à Lei de Acesso à Informação que em seu artigo 31, trouxe regras específicas para proteção de dados pessoais (BRASIL, 2011b).

A Lei nº 13.709/18 surge como uma Lei Geral de Proteção de Dados pessoais (LGPD), possibilitando a integração da proteção à personalidade de forma mais adequada (BRASIL, 2018).

Hodiernamente, a proteção de dados pessoais está presente em mais de 120 países, inclusive no Brasil, conforme apresentado na figura 1:

Figura 1 - Lei de Proteção de dados no mundo



Fonte: CNIL (2018, *online*).

Editada no contexto do Regulamento Geral de Proteção de dados da União Europeia, a LGPD tem por objetivo proteger os direitos fundamentais de liberdade, privacidade e o livre desenvolvimento da personalidade da pessoa natural. A referida lei visa regular o tratamento de dados pessoais compreendendo este, como toda operação realizada com dados desde a coleta, perpassando pelo acesso, uso, transmissão, processamento, armazenamen-to e eliminação, conforme previsão do artigo 5º, inciso X, CF (BRASIL, 1998). Essa lei, busca ainda estabelecer regras sobre o tratamento de dados pessoais pelas empresas e pelo poder público, fixando limites para tal.

O dado pessoal constitui aquelas informações relacionadas à pessoa natural, tendo como fundamentos a previsão do artigo 2º (BRASIL, 2018), que envolvem entre outros, o respeito à privacidade; a autodeterminação informativa; a liberdade de expressão, de informação, de comunicação e de opinião.

Os agentes de tratamento dos dados pessoais são definidos pela lei em seu artigo 5º, ao conceituar as figuras do controlador, operador e do encarregado (BRASIL, 2018). Por controlador, conforme exposto no artigo 5º, inciso VI, compreende-se como "[...] pessoa natural ou jurídica, de direito público ou privado, a quem competem as decisões referentes ao tratamento de dados pessoais". O operador, conforme o inciso VII, do mesmo artigo, compreende a "[...] pessoa natural ou jurídica, de direito público ou privado, que realiza o tratamento de dados pessoais em nome do controlador". E, a figura do encarregado, prevista no inciso VII, constituída pela "[...] pessoa natural, indicada pelo controlador, que atua como canal de comunicação entre o controlador e os titulares e a autoridade nacional" (BRASIL, 2018, *online*).

A referida lei apresenta como princípios que devem ser aplicados no tratamento dos dados pessoais a previsão do artigo 6º, estabelecendo pelo princípio da finalidade, que o referido tratamento deve obedecer a propósitos legítimos, específicos e informados ao titular dos dados; o princípio da necessidade exige que o tratamento esteja limitado à abrangência de dados pertinentes, proporcionais e não excessivos em relação às finalidades propostas; o princípio da qualidade dos dados está atrelado à garantia de exatidão, clareza, relevância e atualização dos dados ao seu titular; a transparência também se trata de um princípio que garante o acesso a informações claras, precisas e de forma fácil.

Pode-se destacar que os contornos objetivos da LGPD, pressupõem o crescimento exponencial da dimensão coletiva da privacidade em que qualquer cidadão está exposto a danos em sua privacidade pelo uso ilícito de dados pessoais.

4.2 Proteção da privacidade e acesso à informação

A utilização de conceitos que envolvem dados e informações exige uma análise sobre a definição dos mesmos devendo ser compreendidos em suas particularidades.

Segundo Le Coadic (1996), os dados consistem numa forma de representação composta da informação codificada, que permite colocá-las sobre o processamento eletrônico. Os dados podem ser considerados como um conjunto de fatos distintos e objetivos, quantificáveis e qualificáveis, em que sozinhos e independentes não conduzem a nenhuma interpretação.

Quanto à informação, a sua criação depende da compreensão dos dados e da atribuição de significado aos mesmos. Segundo Capurro (2003) não há uma definição única sobre o seu significado,

sendo assim de caráter polissêmico. Essa polissemia, permite compreendê-la pelo viés de noção básica, que segundo Capurro e Hjørland (2007), a noção de que se deve ter em mente é que a informação corresponde ao que é informativo para uma determinada pessoa, e o que é informativo depende das necessidades interpretativas e habilidades do indivíduo. Dessa forma, Machlup e Mansfield (1983) já assinalavam que a informação envolve indivíduos transmitindo e recebendo mensagens no contexto de suas ações possíveis.

Estabelecida a diferença entre dado e informação, torna-se necessário evidenciar que as tecnologias da informação e comunicação, têm possibilitado a disponibilidade de dados pessoais em grande volume, o que reforça a vulnerabilidade dos cidadãos frente à violação dos seus direitos fundamentais de liberdade, privacidade e do desenvolvimento da pessoa natural.

Praticamente, em toda atividade realizada pelos indivíduos há produção de dados, mesmo que de forma indireta. Os dados pessoais são considerados atualmente como o principal insumo da economia, pois podem ser tratados e valorados economicamente.

Atualmente, tem se tornado frequente os casos de vazamento de dados por meio de redes sociais, sendo mais evidente o caso do *Facebook*, em que houve vazamento de dados de 87 milhões de usuários pela empresa Cambridge Analítica (FOLHA DE SÃO PAULO, 2018).

O que ocorre na prática trata-se de uma assimetria informacional em que o aumento da quantidade de dados reduz o conhecimento dos cidadãos principalmente sobre o uso de seus próprios dados.

Nesse caso, a LGPD surge para favorecer o controle do cidadão sobre seus dados. Pode-se dizer que a LGPD abandona a ideia do sigilo ao acesso à informação, pois este constitui paradigma de exclusão. Já a proteção de dados, embora pareça paradoxal, tem como uma de suas principais funções proporcionar segurança para que informações pessoais circulem adequada-mente. Para isso, torna o uso do dado pessoal legal, ao buscar estabelecer várias instâncias de controle de forma responsável e tutelada, proporcionando meios claros e seguros para a sua proteção.

Nesse sentido, a proteção da privacidade fica ligada ao controle sobre os dados pessoais, destacando-se o dever de informar, por parte das empresas que lidam com os dados, de forma clara e transparente, sobre o que fazem com os mesmos, como um direito do titular. Por outro lado garante a autodeterminação informativa, como forma do titular decidir o que, quando e como devem ser utilizados os seus dados.

4.3 O consentimento no tratamento de dados pessoais

Como requisito considerado mais importante para o tratamento dos dados pessoais está o consentimento do titular, previsto no inciso I do artigo 7º (BRASIL, 2018). O consentimento implica na permissão do titular dos dados para que alguém possa coletá-los e tratá-los.

De acordo com o artigo 5º, inciso XII, da LGPD, o consentimento constitui manifestação livre, informada e inequívoca, pela qual o titular concorda com o tratamento de seus dados para determinada finalidade. O referido conceito de consentimento no âmbito da LGPD inova ao utilizar o termo inequívoco, ao referir-se a possibilidade de que o mesmo possa ser auditável, uma vez que, segundo Vancim e Neves (2015), no âmbito do Marco Civil da Internet, exige-se que o consentimento seja livre, expresso e informado, conforme exposto em seu inciso VII (BRASIL, 2014).

No artigo 7º da LGPD determina-se que o consentimento deverá ser fornecido por escrito ou por outro meio que demonstre a manifestação inequívoca de vontade do titular. Além disso, a LGPD exige que caso o fornecimento seja realizado de forma escrita, deverá ser realizada por meio de cláusula destacada das demais cláusulas contratuais.

Caso exista alguma dúvida sobre a realização do consentimento por parte do titular, cabe ao controlador o ônus da prova de que o consentimento foi obtido em conformidade com a Lei.

No caso de vício de consentimento, o tratamento de dados pessoais resta vedado, sendo considerado nulo caso as informações fornecidas ao titular tenham conteúdo enganoso ou abusivo, e não se apresente de forma clara e inequívoca. As autorizações genéricas para o tratamento de dados pessoais também são consideradas nulas.

No caso de mudança na finalidade do consentimento não compatível com o original, o controlador deverá informar previamente ao titular sobre as mudanças de finalidade. Nesse caso, o titular poderá revogar o consentimento a qualquer momento mediante manifestação expressa, de forma gratuita e de modo facilitado caso discorde das alterações.

No caso de compartilhamento dos dados pessoais com outros controladores deverá obter consentimento específico do titular para esse fim.

Pode-se dizer que o controle por meio do consentimento do titular, torna-se instrumento de garantia dos direitos de liberdade, intimidade e privacidade, uma vez que o consentimento possibilita a mudança de eixo da estrutura da privacidade constituída pelo cidadão, informação e segredo, para o eixo da tríade cidadão, informação e controle.

A LGPD além de possibilitar a proteção da privacidade no uso da tecnologia, tem como uma de suas principais funções proporcionar segurança para que informações pessoais possam ser devidamente disseminadas. Nesse sentido, o consentimento, torna-se o aspecto mais proeminente para o tratamento dos dados pessoais, pois a circulação sobre o processamento, transmissão e disseminação dos dados e informações dependem da vontade do titular.

O direito à liberdade, à privacidade e ao livre desenvolvimento da pessoa natural, assume novos contornos. Segundo Rodotá (2008, p. 15), a privacidade deve ser considerada também como o direito do cidadão de manter o controle sobre suas próprias informações e de como deve ser definida sua própria esfera particular. Nesse sentido, a ideia passa a ser fortalecida pela LGPD, pois, possuem como fundamento a autodeterminação informativa, que permite ao cidadão definir como seus dados e informações devem ser tratados e disseminados.

5 CONSIDERAÇÕES FINAIS

A discussão sobre proteção de dados pessoais não se considera nova. Vários países do mundo já possuem uma lei efetiva de proteção de dados ou estão em processo de implementação.

No Brasil, a LGPD surge num momento em que a atenção está voltada para a efetivação de garantias constitucionais de liberdade, privacidade e desenvolvimento da pessoa humana. Apesar da existência de várias normas que tratam da proteção de dados, surge a necessidade de estabelecer o diálogo das fontes para que os referidos institutos sejam alinhados e a LGPD possa ter efetividade prática.

Restou claro que a proteção da privacidade também tem como uma de suas principais funções proporcionar segurança para que informações pessoais possam ser disseminadas adequadamente.

O papel do consentimento torna-se o aspecto mais relevante para o tratamento dos dados pessoais, pois a circulação sobre o processamento, transmissão e disseminação dos dados e informações do titular dependem de sua vontade.

Dessa forma, o acesso à informação por meio do controle da LGPD permite a privacidade se manter como ponto central para a proteção dos direitos fundamentais dos indivíduos, indicando que o controle torna-se instrumento de garantia dos direitos de liberdade, intimidade e privacidade.

Nesse sentido, os objetivos propostos neste trabalho foram alcançados por ter conseguido apresentar uma visão sobre como a privacidade e o acesso à informação coadunam para a efetivação das garantias constitucionais.

REFERÊNCIAS

BRASIL. Constituição (1988). **Constituição da República Federativa do Brasil**. Brasília, DF: Senado Federal: Centro Gráfico, 1988, 292 p.

BRASIL. **Lei nº 8.078**, de 11 de setembro de 1990. Dispõe sobre a proteção do consumidor e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/LEIS/L8078.htm>. Acesso em: 6 out. 2018.

BRASIL. **Lei nº 9.507**, de 12 de novembro de 1997. Regula o direito de acesso a informações e disciplina o rito processual do *habeas data*. Disponível em: <http://www.planalto.gov.br/ccivil_03/LEIS/L9507.htm>. Acesso em: 6 out. 2018.

BRASIL. **Lei nº 12.414**, de 9 de junho de 2011. 2011a. Disciplina a formação e consulta a bancos de dados com informações de adimplemento, de pessoas naturais ou de pessoas jurídicas, para formação de histórico de crédito. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/L12414.htm>. Acesso em: 6 out. 2018.

BRASIL. **Lei nº 12.527**, de 18 de novembro de 2011. 2011b. 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: 6 out. 2018.

BRASIL. **Lei nº 12.965**, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm>. Acesso em: 6 out. 2018.

BRASIL. **Lei nº 13.709**, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>. Acesso em: 6 out. 2018.

CAPURRO, R. Epistemologia e Ciência da Informação. In: Encontro Nacional de Pesquisa em Ciência da Informação – Enancib, 5., 2003, Belo Horizonte. **Anais...** Belo Horizonte: ECI/UFMG, 2003.

CAPURRO, R; HJØRLAND, B. O conceito de informação. **Perspectivas em Ciência da Informação**. Belo Horizonte, v. 12, n. 1, p. 148-207, jan./abr. 2007.

CNIL. Protéger les données personnelles, accompagner l'innovation, préserver les libertés individuelles. Data protection around the world. Disponível em: <https://www.cnil.fr/en/data-protection-around-the-world>. Acesso em: 6 out. 2018.

FOLHA DE SÃO PAULO. Facebook e Cambridge Analytica são alvos de ação coletiva. 10 abr. 2018. Disponível em: <https://www1.folha.uol.com.br/mundo/2018/04/facebook-e-cambridge-analytica-sao-alvos-de-acao-coletiva-diz-jornal.shtml>. Acesso em: 8 out. 2018.

HENRIQUES, A.; MEDEIROS. J. B. **Metodologia Científica na Pesquisa Jurídicas**. 9. ed. São Paulo: Atlas, 2017.

LE COADIC, Y. **A ciência da informação**. Brasília, DF: Briquet de Lemos/Livros, 1996.

MACHLUP, F.; MANSFIELD, U. (Ed.). **The study of information: Interdisciplinary** messages. New York, NY: Wiley, 1983.

RODOTÁ, S. **A Vida na Sociedade da Vigilância**: A privacidade hoje. Rio de Janeiro: Renovar, 2008.

SEVERINO, A. J. **Metodologia do Trabalho Científico**. 23. ed. São Paulo: Cortez, 2007.

VANCIM, A. R.; NEVES, F. F. **Marco Civil da Internet**: Anotações à Lei Nº 12.965/2014. São Paulo: Mundo Jurídico, 2015.

LETREIROS E AMORES:
uma estratégia informacional para existir no *Instagram*

SIGNS AND LOVE:
an informational strategy to exist on Instagram

Matheus José Pessoa de Andrade[1], Tassyara Onofre de Oliveira[2]
(1) Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba – PPGCI/ UFPB, theujp@hotmail.com.
(2) Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba – PPGCI/ UFPB, tassyaraonofre@gmail.com.

Resumo
O presente trabalho tem como objetivo pensar a função informacional dos letreiros de cidade. Acreditamos, a priori, ser uma estratégia de existência para a internet, atendendo as necessidades e valores contidos nas informações das redes sociais digitais. Desta maneira, partimos do pressuposto que o Instagram modificou os ambientes sociais pelas formas e fluxos de compartilhamento das práticas fotográficas, fazendo com que as coisas se tornem instagramáveis, sobretudo no universo dos viajantes fotógrafos. Para atender a esse regime informacional, proliferam-se os letreiros turísticos nas cidades. No caso específico do letreiro da cidade de João Pessoa, o qual nós abordamos, foi realizada uma análise de dados para averiguar o número de postagens e acessos baseado na hastag #euamojampa para, juntamente com nossas reflexões teóricas, atestar a função informacional dos letreiros de cidade em tempos de vida online.
Palavras-chave: Letreiros de cidade. Informação. Instagram. Fotografia.

Abstract
The present work aims to think the informational function of the city signs. We believe, a priori, to be a strategy of existence for the internet, meeting the needs and values contained in the information of digital social networks. In this way, we start from the assumption that the Instagram modified the social environments by the forms and streams of sharing of the photographic practices, making things become unstable, especially in the universe of photographic travelers. In order to meet this informational regime, tourist signs are proliferating in the cities. In the specific case of the sign of the city of João Pessoa, which we approached, a data analysis was carried out to ascertain the number of posts and accesses based on the #euamojampa hastag to, together with our theoretical reflections, to verify the informational function of the signs of city in times of life online.
Keywords: City signs. Information. Instagram. Photography

1 CONTEXTUALIZAÇÃO INICIAL

No século XXI, foram construídos vários letreiros nas mais diversas cidades do Brasil e do mundo, os quais nitidamente fazem parte da atividade econômica do turismo. Muitas vezes, os letreiros são esteticamente bem feitos, coloridos, em paisagens privilegiadas, insinuando o amor instantâneo. Algumas vezes estão localizados em pontos estratégicos da cidade, noutros casos são construídos num novo lugar, criando novas práticas de sociabilidades.

O turista, também, é um colecionador fotografias, como um tipo de suvenir, um registro dos momentos únicos que o viajante vivencia em lugares turisticamente importantes. Esses momentos migraram das estantes de casa, dos quadros e porta retratos, para as telas de computador e dispositivos móveis, possibilitando uma maior visibilidade a essa experiência. Nesse campo, os letreiros tornaram-se um lugar em potencial de fotografia turísticas.

O fluxo de informações através da fotografia digital na internet passa a ter um papel abertamente indispensável, visto que a sociedade vivencia um novo regime de conexão através da internet, fruto das necessidades de troca de informação entre as pessoas. Nesse contexto, redes sociais como o *Instagram* podem ser vistas como vias de compartilhamento de informações em tempo real, considerando a dinâmica do desenvolvimento tecnológico. Devido à quantidade de usuários ativos existentes e suas câmeras portáteis, a informação visual consegue ser montada e disseminada instantaneamente.

2 OBJETIVO NORTEADOR

Nossa proposta de análise dos letreiros de cidade está diretamente relacionada ao *Instagram*. Considerando o letreiro como uma estratégia de existir *online*, reconhecemos o impacto causado pelo *Instagram* nesse contexto. Pensamos, assim, na seguinte questão: qual a função informacional dos letreiros de cidade em tempo de *Instagram*?

3 CARACTERIZAÇÃO DA PESQUISA

Considerando o objetivo proposto, foi desenvolvida uma pesquisa com abordagem de cunho bibliográfico, no tocante a analise de traçar um histórico sobre o objeto de estudo, como também ajuda a identificar contradições e respostas sobre o objeto. Bem como uma investigação exploratória e descritiva, no que concerne a analise de dados coletados sobre a tag #euamojampa através da ferramenta *Hashtracking*.

4 INSTAGRAM: POR UMA EXISTÊNCIA ONLINE

Falar sobre redes sociais é apontar para um lugar de sociabilidades, assim como são as praças ou os bares, por exemplo. Ou como um dia foram algumas calçadas das casas em pequenas

cidades. Nelas surgem valores inerentes daqueles lugares de encontro e daquelas formas de relações sociais que ali se estabelecem. Práticas de horários, de conduta, de presença e ausência, e até de tipo de informação.

Na era das redes sociais digitais criou-se um lugar de vivências *online*, onde nos incluímos também pela possibilidade de participar e comprovar diariamente nossa existência social. Seja qual for a maneira que utilizam, os usuários buscam também afeto, pertencimento, reconhecimento, amor e felicidade. Procuram, a seu modo, encontrar satisfação para que seus desejos sejam realizados, como também, vencer a solidão, o tédio, fugir da realidade e até mesmo tentar seus "minutos de fama". Enquanto uns chegam à fama e ganham grande notoriedade, outros fazem o possível para prezar por sua transparência. Trata-se, portanto, da percepção de uma prática relacional construída naquele lugar, onde se luta pela existência em rede e pelo ato de ser curtido, e até idolatrado pelos outros participantes.

É cabível observar que existimos nas redes sociais digitais também por se dar conta de um tipo de valor simbólico ali impregnado, como Márcia Tiburi destaca: "Não aparecemos apenas porque esse é nosso modo de habitar a cidade, a sociedade, de estar no mundo. Aparecer tornou-se urgente, como forma de capital" (2017, p.121). Por isso, postam-se diariamente informações sobre as coisas do mundo. Conteúdos diversos, em formatos de textos, imagens, áudios e vídeos, de acordo com as necessidades de cada integrante da rede social *Instagram*.

Entendemos que ele é uma rede social predominantemente de fotografias, diferentes de outras tantas. Linascheke (2011) ressalta que o *Instagram* é uma comunidade fotográfica, e isso explica tamanho sucesso. Comparando-o com redes sociais como o *Twitter* e o *Facebook*, o autor observa que o *Instagram* é a única rede social inteiramente baseada em fotografias. Há algum texto permeando tudo isso, alguns comentários, e outras tantas curtidas, mas mesmo assim, segundo ele, tudo se resume à imagem fotográfica e é justamente por isso que o *Instagram* é único, não há como fazer uma postagem sem se utilizar de uma imagem.

Nada mais atual para se comparar com a obra *A sociedade do espetáculo* de Guy Debord do que a rede social *Instagram*. O espetáculo, de acordo com o pensamento debordiano, não é apenas um conjunto de imagens, mas uma relação social entre pessoas, mediatizada por imagens.

Com base na reflexão de Susan Sontag, desde a invenção da fotografia, no século XIX, o mundo vivencia uma mudança no paradigma das visualidades, pelas práticas, valores e costumes circunscritos no uso da câmera fotográfica nos mais diversos grupos sociais. Os desdobramentos dados pela fotografia foram capazes de transformar a sociedade. Por exemplo, "viajar se torna uma estratégia de acumular fotos" (SONTAG, 2004, p.20), e não apenas de desbravar lugares, conhecer pessoas e relatar em textos.

A interferência do ato fotográfico sobre as coisas do mundo pode ser percebida em pontos turísticos, eventos, fachadas, vestimentas etc., cujas estruturas muitas vezes são feitas para serem fotografadas, isto é, para existir no campo visual das fotos. Para Sontag, em síntese, "hoje, tudo existe para terminar numa fotografia" (2004, p.35). Para nós, o *Instagram* potencializou esse aforismo principalmente pelo sistema de circulação das imagens, amplificando os tratamentos visuais das coisas para serem fotografadas e postadas potencialmente.

Em consonância com os aparelhos celulares, pela portabilidade das câmeras e sua conexão com as redes sociais virtuais, podemos atestar dois aspectos da cultura fotográfica no *Instagram*: de

um lado, pessoas capturando tudo que convém para partilhar com os outros nas redes sociais a ideia de que estiveram em determinado lugar; e, do outro lado, os lugares mendigando cada vez mais para existirem dentro dos jogos simbólicos de fotografia instaurados no cotidiano das redes sociais virtuais, pelas pessoas que capturam incessantemente suas fotos. Por isso, os lugares buscam estratégias, tais como: os letreiros para participarem deste regime. Eles são potencialmente fotografáveis e, por tabela, fazem o lugar existir no fluxo informativo dessas imagens colecionadas pelos viajantes, e publicamente postas numa plataforma como o *Instagram*.

Com o auxilio das *hastags*, muitas vezes já induzidas pelos próprios monumentos , os usuários criam "rastros" que são publicados, arquivados, e, portanto, são recuperáveis e buscáveis. Isso resulta numa visibilidade imediata e "gratuita" para as cidades que possuem tais letreiros. Essa funcionalidade locativa abriu caminho para um compartilhamento de *insights* do cotidiano das pessoas em tempo real.

Os letreiros de cidades, construídos nesse modismo da segunda década do século XXI, foram impulsionados pelas lógicas simbólicas das redes sociais digitais, sobretudo na prática fotográfica dos viajantes usuários do Instagram. Eles são feitos para fazer com que as cidades participem enquanto lugar nesse fluxo de visibilidade na internet, aumentando assim o seu capital simbólico. São monumentos ao *self*, por assim dizer, feitos para terminarem numa fotografia, aparecerem online e se tornarem *hastags*. Buscam, assim, se fortalecer enquanto forma de valor informacional no *Instagram*, contribuindo com o espetáculo para cumprir com o desdobramento contido nos interesses dos gestores.

4 LETREIROS DE CIDADE: *A INSTAGRAMIZAÇÃO* DAS COISAS

Para entender melhor o encantamento pelos letreiros e compreender por que eles atraem tantos olhares, câmeras fotográficas e corações, basta voltar no tempo até 1923, ano de inauguração do símbolo em *Hollywood*. A princípio a intenção da construção não foi tornar a cidade famosa. O motivo de sua criação foi como peça publicitária para vender lotes do empreendimento imobiliário. A proposta era deixar as letras posicionadas ali durante um ano e meio para incentivar a compra de terrenos próximos. Por isso, a primeira aparição das letras continha, além do nome da cidade, o complemento "land", que significa "terra" em inglês, formando a palavra *Hollywoodland*. O que inicialmente era uma ação de publicidade caiu na graça dos moradores e foi parar nas telas de cinema do mundo, acompanhando a ascensão do cinema americano.

No continente europeu, o slogan *I Amsterdam* faz um trocadilho na língua inglesa, sugerindo a frase "Eu sou Amsterdã". Localizado na parte de trás do *Rijksmuseumque* e que é o museu de arte mais importante da Holanda, na *Museumplein*, o monumento tornou-se rapidamente um ícone da cidade e uma das fotografias mais procuradas para quem visita Amsterdam, a capital holandesa.

A região norte do Caribe também não deixou a onda passar. Cancún, o paraíso mexicano de ilhas, também tem seu famoso letreiro turístico. O monumento chamado de *Cancun Sign* é um grande e colorido letreiro, aberto oficialmente para visitantes em 31 de julho de 2014, uma das mais famosas atrações para os visitantes e turistas da ilha mexicana fica na famosa *Playa Delfines*,

conhecida por ser a mais bela de Cancún e também chamada de *El Mirador*, por oferecer uma bela vista do lugar.

Os três letreiros citados foram inesgotavelmente fotografados por milhares de turistas desde suas criações, tornando-se lugares obrigatórios de fotografia para viajantes que ali passavam.

Sob a égide da imagem digital, o *Instagram* é um espaço fotográfico inventivo, o qual permite variadas maneiras de utilização dos perfis. Uma forma comum de uso está relacionada com a amostragem do cotidiano de si. Os usuários são um tipo de curadores dos próprios perfis, ao elegerem critérios pessoais de publicação que, normamente, não denunciam contra a projeção de si: lugares, objetos, pessoas e performances devem obedecer a uma lógica naquela forma de exposição para os seus seguidores, isto é, dentro de um regime político de produção de discursos sob vigilância.

As coisas do mundo passaram a serem pensadas para serem fotografadas para postagens e consequentemente existirem nas imagens da rede social virtuais, semeando seu capital simbólico. Aqui consideramos o que Marshall McLuhan explica quando privilegia um olhar sobre o impacto do meio de comunicação na vida social, antes mesmo de se preocupar com os efeitos ideológicos da informação sobre o receptor. Para ele, "(...) é o meio que configura e controla a proporção e a forma das ações e associações humanas" (2005, p.23). Nesse sentido, o meio recria as relações interpessoais e as formas dos ambientes, notado por nós quando as coisas passaram a se orientar visualmente numa conotação sugerida pelo funcionamento do *Instagram*. Por exemplo, estabelecimentos preparam cenários e iluminações para seus clientes fazerem suas *selfs*; restaurantes dão maior atenção ao visual de seus *drinks* para prosperar nos perfis dos clientes; eventos montam lugares para seus visitantes fazerem seus registros fotográficos com *hastag* sugerida, amplificando sua divulgação; enfim, podemos identificar uma gama de gestos em função dessa lógica, inclusive os letreiros de cidade, a fim de se tornarem fotografias e *hastags* na plataforma, dentro dos parâmetros de curadoria dos usuários.

> Como efeito disso, a instagramização dos espaços físicos, públicos e privados, dos objetos, das formas de circulação na cidade e de experiências envolve reordená-los e editá-los para que sejam visíveis e publicáveis (BENTES, 2017).

Trata-se de uma *instagramização* das cidades que ajustam seu ambiente a fim de existir *online*. Muitos dos letreiros de cidade escrevem uma forma de amor de passagem (e não passageiro), mas que, na ocasião, são registrados para perpetuar enquanto memória fotográfica. O amor expresso parece estar dentro dos critérios dos curadores e vigilantes do *Instagram*, principalmente pelas relações estabelecidas das pessoas com os lugares. Ao tempo que constroem uma forma de projetar um lugar afetivamente acolhedor para os turistas (e talvez não seja), fundamentados na estratégia de fortalecer o mercado turístico daquele espaço. Tornam-se, assim, instagramáveis, ao participarem desse jogo.

Concordamos com Bauman (2001) quando explica a sociedade a partir da concepção de liquidez, calcada na cultura do consumo como base de valor de toda sociabilidade dos indivíduos. A vida não se trata apenas de ser, mas também de ter. "(...) A satisfação de alcançar um objeto é apenas momentânea" (2001, p.70). Assim como o consumo, centrado no ato efêmero de obter algo e, logo, passar para o próximo item. Os amores ganham novas significações por parte de uma vida liquefeita e veloz como a atual. O amor duradouro é remodelado nas práticas de sociabilidades

modernas, homogeneizada com os valores culturais do consumo contido nas formas de relações sociais amorosas. Em vias virtuais de relacionamento, "muito mais tem acontecido no caminho em direção à líquida e individualizada sociedade moderna para tornar os compromissos de longo prazo pouco numerosos (...)" (BAUMAN, 2004, p.61). É apenas um amor rápido, de passagem, fixado numa foto e postado para um grupo de seguidores entenderem que alguém esteve naquela cidade, ao tempo que os seguidores conhecem fotograficamente aquele lugar através das diversas informações reconhecíveis – muitas vezes nada se sabe contextualmente daquele acontecimento pela sua forma de representação.

Nesse sentido, a ação informativa do letreiro de cidade se rende às mudanças ocorridas a partir da invenção do *Instagram*, agregando valores como o registro, o amor e a recordação na prática dos fotógrafos viajantes, dentro dessa lógica de existência *online* na plataforma em questão.

5 #EUAMOJAMPA: O CASO DO LETREIRO DA CIDADE DE JOÃO PESSOA

Entre tantos letreiros famosos espalhados pelo mundo, a capital Paraibana, João Pessoa, sentiu a necessidade de possuir também seu monumento e declarar aos seus visitantes o amor pela cidade.

O letreiro com a frase "Eu amo Jampa" ou "Eu <3 Jampa", foi inaugurado em 28 de dezembro de 2015, situado no Busto de Tamandaré, que fica entre as praias de Tambaú e Cabo Branco, ambas consideradas as praias mais famosas da cidade. O letreiro, segundo a gestão municipal, é uma homenagem e representa "o amor pela cidade estampado em um dos principais pontos turísticos da capital". Desde sua fundação o local onde se está o letreiro costuma ser visitado por turistas e pessoenses, um dos cenários mais disputados para fotos e *selfie*s. Um reflexo disso são os números oferecidos pelo *Instragram* de postagens realizadas utilização da localização e a *hashtag*. Em uma breve pesquisa de dados no próprio aplicativo verificamos o número de 41.512 publicações públicas até o dia 23 de agosto de 2018.

Diante dessa expressiva quantidade de publicações, resolvemos fazer um diagnóstico mais preciso para entender melhor a dimensão compartilhamentos de informações fotográficas relacionadas ao letreiro. Para a análise foi escolhida a *hashtag* #euamojampa que faz relação ao próprio letreiro e a ferramenta de rastreio foi a *Hashtracking* em publicações públicas, que oferece o serviço de rastreamento de *hashtag* no *Instagram*. O diagnóstico foi realizado entre os dias 16 a 23 de agosto de 2018, exatamente durante uma semana. Os seguintes dados foram apresentados, conforme as figuras 01 e 02 (verificar figuras no apêndice A). A Figura 01 apresenta que, durante uma semana, os usuários do *Instagram*, utilizando a *hashtag* #euamojampa fizeram 349 postagens, o que gerou 631 comentários e 17.906 curtidas, 3.054,545 impressões (número de vezes que seu conteúdo é exibido, independentemente de ter sido clicado ou não) e o alcance de 1.003,57 (número total de pessoas que visualizam seu conteúdo). Já a figura 05 expõe como funciona o engajamento em relação à #euamojampa no *Instagram*.

Desta amostragem mínima, de uma semana, acreditamos ter um número significativo de visualizações a ponto de fazer com que as cidades e os demais lugares criem tipos de estratégia, como os letreiros, para produzirem informação fotográfica e passarem a existirem com veemência no ambiente do *Instagram*.

A figura 03 (verificar figura no apêndice A) apresenta as *buzzywords* relacionadas à *hashtag* #euamojampa. Foram encontradas 46 palavras que foram utilizadas em postagens junto *tag* analisada. Isso comprova que a repercussão de uma imagem fotográfica feita no letreiro da cidade e postada no Instagram pode adquirir uma dimensão de visibilidade incalculável, dando subsídio à importância do monumento enquanto proliferador de informação nas redes sociais digitais. A figura 04 (verificar figura no apêndice A) aponta a geolocalização de postagens realizadas com a *hashtag* em questão. A ferramenta verificou 16 locais distintos, todos no Brasil, apenas numa semana de análise de dados.

Levamos em consideração também uma série de *hastags* correlatas capazes de representar o letreiro em questão (#jampa, #eus2jampa, #euamojp), de perfis pessoais que, às vezes, não colocam as tags em suas postagens e perfis profissionais que replicam imagens e buscam emplacar outras *hastags* com esse tipo de informação.

Desse vasto número de publicações, acreditamos que essa produção informativa expressiva atesta a importância do letreiro na função de fazer a cidade circular nas plataformas virtuais. A partir da pluralidade de versões fotográficas dos turistas que passam por João Pessoa, ancoradas no letreiro da cidade, cumprem o papel de existir no *Instagram*. Aliás, consideramos que a criação dos letreiros de cidade são impulsionas pela nova mídia que se instaurou. Imaginamos, por comparação, os letreiros mais famosos, das cidades que realmente possuem um mercado turístico expressivo mundialmente.

5 CONSIDERAÇÕES FINAIS

Diante do exposto, acreditamos que os letreiros de cidade são estratégias de informação eficientes para que os lugares existam *online* no *Instagram*, atendendo à lógica do capital simbólico que ali se constitui. Através da informação fotográfica praticada pelos turistas, os lugares se propagam para o mundo, divulgando uma forma de amor simbólico a fim de atender aos interesses mais diversos.

O *Instagram* deve ser considerado, portanto, como um meio de comunicação de significativo impacto social, a ponto de percebermos sua intervenção cultural, tanto nos mais diversos ambientes sociais, os quais frequentamos diariamente, quanto nas relações interpessoais que, através de estratégias, almejam repercutir na rede social virtual em busca de existir no fluxo da sociedade e construir seu capital simbólico através dos *likes*, comentários e *reposts* da vida.

REFERÊNCIAS

BAUMAN, Zygmund. **Modernidade líquida**. Rio de Janeiro: Jorge Zahar Editor, 2001.

_____. **Amor líquido**: sobre a fragilidade dos laços humanos. Rio de Janeiro: Jorge Zahar Editor, 2004.

BENTES, Anna. Instagramização da vida: uma curadoria do visível. In: **Dobras**, UFRJ, 2017.

DEBORD, Guy. **A sociedade do espetáculo**. 11ed. Rio de Janeiro: Contraponto, 2009.

LINASCHEKE, J. **Getting the most from Instagram**. Berkeley: Peachpit Press. 2011.

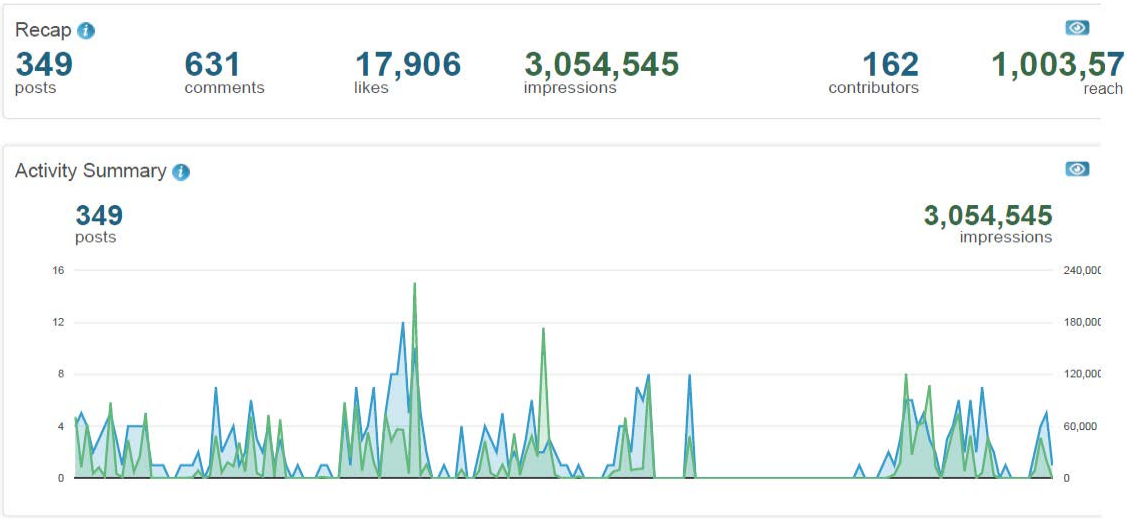
MCLUHAN, Marshall. **Os meios de comunicação como extensões do homem**. 20ed. São Paulo: Cultrix, 2005.

SONTAG, Susan. **Sobre fotografia**. 5ed. São Paulo: Companhia das letras, 2004.

TIBURI, Marcia. Pós-verdade, pós-ética. In: Ética e pós-verdade. Porto Alegre: Dunblinense, 2017.

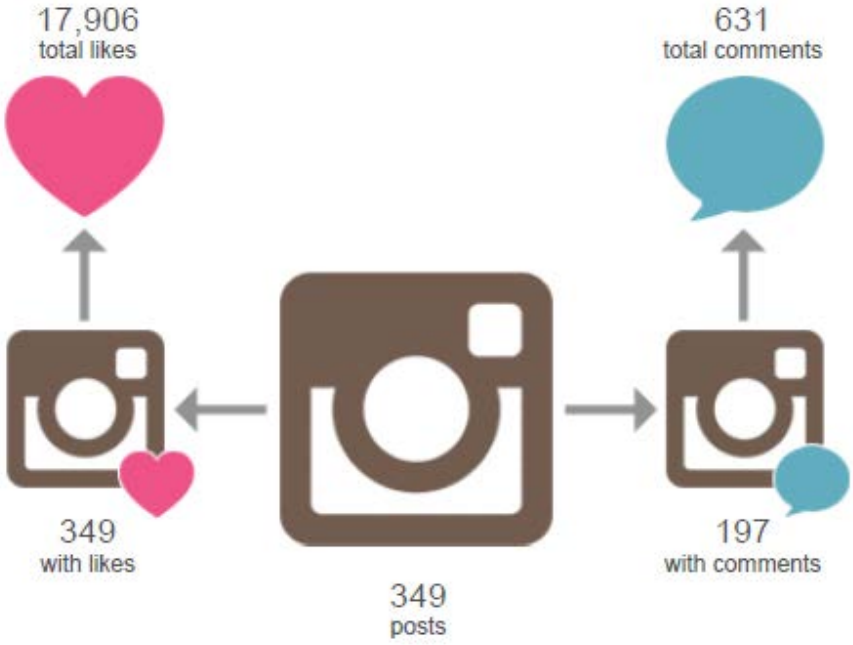
APÊNDICE A

Figura 01: Activity #euamojampa in 8/16/2018 11:05 am — 8/23/2018 10:25.



Fonte: Ha*shtracking*, 2018.

Figura 02: Engagement Rate #euamojampa in 8/16/2018 11:05 am — 8/23/2018 10:25.



Fonte: Hashtracking, 2018.

Figura 03: Buzz Words subjects #euamojampa in 8/16/2018 11:05 am — 8/23/2018 10:25.

Fonte: Hashtracking, 2018.

Figura 04: Geolocation of posts #euamojampa in 8/16/2018 11:05 am — 8/23/2018 10:25.

Fonte: Hashtracking, 2018.

METADADOS E TWITTER:
uso do Tweet Object para identificação de local

METADATA AND TWITTER:
use of Tweet Object for location identification

Denysson Axel Ribeiro Mota[1], Gracy Kelli Martins[(2)]
(1) Universidade Federal do Cariri (UFCA), Av. Tenente Raimundo Rocha S/N - Bairro Cidade Universitária - Juazeiro do Norte - CE – Brasil, CEP 63048-080, denysson.mota@ufca.edu.br
(2) Universidade Federal da Paraíba (UFPB), Cidade Universitária - João Pessoa - PB – Brasil, CEP: 58051-900, gracykmg@gmail.com

Resumo:
Esta pesquisa apresenta um estudo, com base nos Estudos Métricos da Ciência da Informação, sobre a disseminação de informações na plataforma Twitter. Este trabalho teve como objetivos identificar elementos para análises métricas nos Tweets, utilizando como recorte mensagens coletadas na ferramenta via API. Para alcançar estes objetivos, utilizou dados obtidos via API do Twitter, usando sistema operacional Debian Linux e uma aplicação Ruby, armazenados em formato JSON em uma base de dados no SGBD PostGreSQL, e manipulados no SublimeText 3 e OpenOffice quando necessário. Apresenta como resultados a distribuição geográfica em cada uma das etiquetas analisadas, mostrando que mesmo que existem diferenças em valores totais, a distribuição em geral se mantém parecida. Conclui-se que as etiquetas disponíveis no JSON são extremamente úteis, mas que é necessário mais estudo sobre as possíveis formas de identificação de metadados úteis para os processos métricos de informação.
Palavras-chave: Disseminação de dados. Geolocalização. Tweet Object. Twitter.

Abstract:
This research proposes to study, based on the Metric Studies of Information Science, the dissemination of information on the Twitter platform. This work aimed to identify elements for metric analysis in Tweets, using as clipping messages collected in the tool via API. To achieve these goals, it used data obtained via the Twitter API, using Debian Linux operating system and a Ruby application, stored in JSON format in a database in PostGreSQL DBMS, and manipulated in SublimeText 3 and OpenOffice when required. It presents as results the geographical distribution in each of the analyzed tags, showing that even if there are differences in total values, the distribution in general remains similar. It is concluded that the tags available in JSON are extremely useful, but that more study is needed on possible ways of identifying metadata useful for metric information processes.
Keywords: Data dissemination. Geolocation. Tweet Object. Twitter.

1 INTRODUÇÃO

Em momentos de grande comoção, como desastres naturais, surtos epidêmicos e eleições, por exemplo, a sociedade busca por informações relevantes para guiar suas decisões no dia a dia, sejam em rádios, jornais e programas televisivos, e em especial pela internet, em portais de notícias, serviços de vídeo, blogs e redes sociais.

As redes sociais em ambiência digital são ferramentas de comunicação que surgiram no final do século XX, iniciando em 2003 com Myspace, que buscava integrar artistas com fãs, e LinkedIn, criada especificamente como uma rede social de negócios. Posteriormente, em 2004, surgiram duas redes: Orkut, baseado na teoria do mundo pequeno, também conhecida como a teoria dos seis graus de separação, de Stanley Milgram (1967), e *Facebook*, atualmente a rede social mais utilizada, com cerca de 2.2 bilhões de usuários mensais (STATISTA, 2018a). Em 2006 surge o *Twitter*, que dentre estes últimos, é um serviço que se caracterizou pela rápida velocidade de disseminação de informações.

O *Twitter*, desde sua criação em 2006, tem passado por diversas mudanças ao longo de sua história. No entanto algo que se mantém constante é o seu poder como ferramenta de comunicação, principalmente devido à sua estrutura de organização em formato de seguidores, em vez de contatos e/ou amigos como as demais. Criado como uma ferramenta para simular um serviço de SMS na Web, passando para a estrutura de *microblog*, e evoluindo para ferramenta de rede social, o *Twitter* conta atualmente com 330 milhões de usuários ativos mensais em todo o mundo (TWITTER, 2018; STATISTA, 2018b), e por ela são enviadas cerca de 500 milhões de novas mensagens por dia (OMNICORE, 2018), entre estas mensagens informativas enviadas por diversos perfis governamentais oficiais.

Todo este fluxo de mensagens não é composto apenas pelos textos enviados. Junto com a mensagem, o *Twitter* armazena alguns metadados que permitem identificar além das mensagens e outras características, por meio de uma estrutura que permite sua consulta posterior, chamada de *Tweet Object* (TWITTER, 2017a). O *Tweet Object*, que será chamado daqui em diante de TO, contém em sua estrutura dados como a identificação do usuário que realizou o *tweet*, dia e hora do *tweet*, se é uma mensagem original ou uma replicação (chamado de *retweet*), se outros usuários foram citados, entre outros.

Esse tráfego de dados pode ser acessado via APIdo *Twitter*, que fornece acesso à base de dados retornando a solicitação via documento texto em formato *JavaScript Object Notation* (TWITTER, 2017b). A *JavaScript Object Notation*, mas mais conhecido pelo seu acrônimo: JSON, é um padrão de notação para intercâmbio de dados estruturados na internet via texto, criado em 2001 e adotado em diversas plataformas devido a ser independente de linguagem, suas definições podem ser encontradas na RFC8259 e ECMA-404 (BRAY, 2017; ECMA, 2017).

2 OBJETIVOS

Este trabalho tem como objetivo principal identificar elementos para análises métricas nos *tweets*, utilizando como recorte mensagens coletadas na ferramenta via API. Como objetivos específicos, temos: a) identificação dos elementos no *Tweet Object*; b) manipulação dos objetos JSONB na base de dados; c) exportação dos dados para ferramentas de mapas *online*.

3 PROCEDIMENTOS METODOLÓGICOS

A captura dos *tweets* foi realizada via acesso à API usando um *script* na linguagem Ruby, que foi executado em um servidor Debian no período compreendido entre outubro de 2017 até março de 2018, e armazenados em banco de dados PostgreSQL em formato JSON.

Após os estudos da API e da estrutura do *Tweet Object*, o elemento que despertou a curiosidade no primeiro momento a possibilidade de descobrir a distribuição geográfica desses *tweets*.

Para isto inicialmente foram selecionados os *tweets* onde a tag [coordinates] não estava vazia ou nula. Esta etiqueta retorna a localização onde o *tweet* foi realizado, em coordenadas com longitude e latitude, nesta ordem. O Comando SQL, que busca todos os registros de coordenadas que não estiverem vários ou nulos, está exposto na Figura 1.

Figura 1 – Distribuição de Tweets via Etiqueta [coordinates]

```
select
  dados->'coordinates' as coord
from tweets
where
  dados->'coordinates' != '"[]"' and
  dados->'coordinates' != 'null';
```

Fonte: Dados da Pesquisa, 2018.

Após a extração destes dados, foi realizado mapeamento das coordenadas na ferramenta *MapMakerApp*, retornando a o mapa exibido no Apêndice A. Esta plataforma mapa permite a navegação por área, detalhando as quantidades de *tweets* por local, com zoom e navegação interativa.

No entanto apenas 2.125, do total de 2.907.720 de *tweets*, contém a etiqueta [coordinates] preenchida, pois é um campo opcional e recomendado o não preenchimento por organizações de segurança. Foi realizada então uma extração usando a subetiqueta [location] dentro da etiqueta [user]. Esta etiqueta também é opcional, mas está preenchida em 2.003.338 dos *tweets* coletados.

Utilizando o comando exposto na Figura 2, foram obtidos 2.003.338 *tweets* que têm esta etiqueta, que posteriormente foram importados na plataforma Google Maps, que permite a inserção de locais em texto, e mediante a API de Geocoding disponível marca a localização em latitude e longitude no mapa.

Figura 2– Distribuição de Tweets via Etiqueta [user]->[location]

```
select
  dados->'user'->'location' as loc
from tweets
where
  dados->'user'->'location' != '""'
  and
  dados->'user'->'location' != 'null';
```

Fonte: Dados da Pesquisa, 2018.

Os primeiros 12 mil registros foram inseridos no site *Google Maps*, pois ele permite a inserção de locais por nome e a própria plataforma faz a conversão para coordenadas, resultando no mapa exibido aqui no Apêndice B.

No entanto, devido ao grande número de registros, e ao limite de upload da ferramenta, que apenas permite que 2.000 registros sejam lançados por vez, foi decidido seguir por um caminho de marcar apenas uma vez por local.

Considerando que o preenchimento é livre, podendo inclusive conter caracteres especiais, os registros foram então tratados e filtrados usando funcionalidades do *Sublime Text* 3 de busca usando expressões regulares, em conjunto com o pacote *Remove Non Ascii Chars* para remover todos os caracteres especiais que o Twitter permite incluir na localização, como *emoticons*, por exemplo. É importante ressaltar que isto também elimina os registros de cidades que tenham sido redigidos usando ideogramas, como de países da Ásia, mas que devido aos pesquisadores não conhecerem os caracteres não seria possível filtrar o que de fato é cidade e o que não é.

Utilizando a ferramenta de substituição de texto do *Sublime Text* 3, foram utilizadas expressões regulares para excluir, do início e do final da linha, usando os caracteres ^ e $, respectivamente, todos os caracteres não alfanuméricos e espaços em branco ou tabulação. Esta substituição foi realizada de forma repetitiva até que a busca não encontrasse mais registros que atendessem os requisitos.

Posteriormente a esta busca foram excluídas todas as linhas vazias, também utilizando a ferramenta de busca textual e informando a expressão regular ^\n e deixando o campo {Replace:} vazio. Depois disto, foram excluídas as linhas duplicadas no menu {Edit}>{Permute Lines}>{Unique}. Com estes mecanismos, foi possível reduzir o número de registros de locais para 129.875, dos 2.003.338 registros recuperados anteriormente.

Mesmo reduzindo o número total de registros, a limitação da plataforma *Google Maps* ainda é uma barreira para o adequado uso dela, sendo necessária a realização de 65 processos de importação, além das correções e validações necessárias nos casos em que a plataforma não compreenda o local.

Por fim, foram realizadas extrações usando a etiqueta [place]. Esta etiqueta indica a qual local o *tweet* está relacionado, mas não indica, necessariamente, onde ele foi realizado. Dentro desta etiqueta existem duas outras que são interessantes: [full_name] e [bounding_box]. A primeira tem o local em formato texto, passível de importação no *Google Maps*, enquanto a segunda tem uma outra etiqueta aninhada chamada de [coordinates], que contém as coordenadas para formar uma caixa de exibição do local. O comando SQL utilizado pode ser visto na Figura 3.

Figura 3– Distribuição de Tweets via Etiqueta [place]->[bounding_box]->[coordinates]

```
select
  dados->'place'->'bounding_box'->'coordinates' as lugar
from tweets
where
  dados->'place' != '""' and
  dados->'place' != 'null';
```

Fonte: Dados da Pesquisa, 2018.

Devido à limitação de importação do *Google Maps*, decidiu-se usar esta segunda opção, tratando as coordenadas no *OpenOffice* para obter o ponto central indicado. Após a extração destas

coordenadas, as mesmas foram importadas na plataforma *MapMakerApp*, resultando no mapa exibido no Apêndice C.

4 RESULTADOS

É possível afirmar, com a análise apresentada, que os três objetivos foram alcançados de forma total, mas que ainda permitem a expansão da pesquisa com novas etiquetas e análises.

As diferentes possibilidades de coleta de localização dos tweets mostram a flexibilidade da ferramenta. Mesmo com as diferenças entre os valores obtidos, como pode ser visto nos Apêndices A,B e C, os grandes aglomerados de tweet permanecem (Brasil, EUA, India e Tailândia).

Isto pode demonstrar alguma consistência dos metadados coletados automaticamente pela plataforma com os informados livremente pelos usuários, mas mais testes são necessários considerando as possibilidades de representação na etiqueta [user]->[location].

5 CONSIDERAÇÕES FINAIS

A plataforma disponibilizada pelo Twitter nos permite avaliar diversos metadados para estudos métricos da informação. No entanto, é necessária a identificação das etiquetas disponíveis, para saber onde ir para buscar a informação necessária, e quais as possibilidades de combinação entre elas para enriquecimento da análise.

Nesta perspectiva, buscou-se aqui estudar as etiquetas e geolocalização, de forma relativamente simples, mas é possível realizar posteriores pesquisas que envolvam cruzamentos destes dados com outras etiquetas, e com isso obter novas possibilidades de análise.

De forma semelhante, podem ser feitas comparações entre os metadados das etiquetas [coordinates] da raiz do tweet com os da etiqueta [place]->[bounding_box]->[coordinates] dos tweets que tenham ambas.

A possibilidade de importação destes metadados em plataformas de mapas que permitem navegação é um caminho interessante de disponibilização de dados de pesquisa para uso posterior.

REFERÊNCIAS

BRAY, T. (Ed.). The JavaScript Object Notation (JSON) Data Interchange Format. 2017. Disponível em <https://www.rfc-editor.org/info/rfc8259> Acesso em: < 25 mar. 2018 >.

ECMA INTERNATIONAL (ECMA). The JSON Data Interchange Syntax. 2017. Disponível em <https://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf > Acesso em: < 25 mar. 2018>.

MILGRAM, S. The small world problem. **Psychology Today**, v. 2, p. 60-67, 1967.

STATISTA. Most popular social networks worldwide as of April 2018, ranked by number of active users (in millions). 2018a. Disponível em < https://www.statista.com/statistics/272014/global-social--networks-ranked-by-number-of-users/ > Acesso em: < 6 abr. 2018 >.

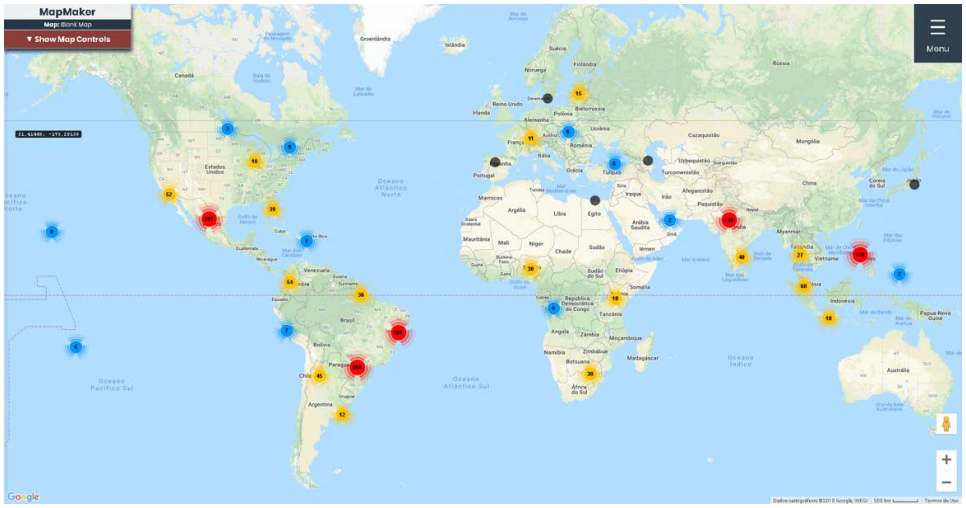
STATISTA. Twitter: number of monthly active users 2010-2017. 2018b. Disponível em <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ > Acesso em: < 6 abr. 2018 >.

TWITTER. Q1 2018 - Letter to Shareholders. 2018. Disponível em <http://files.shareholder.com/downloads/AMDA-2F526X/6236645277x0x978181/2FD6D58F-A930-4EB2-90B0-9C3A120648DE/Q1_2018_Shareholder_Letter.pdf> Acesso em: < 03 maio 2018 >.

TWITTER. Tweet Object – Twitter Developers. 2017a. Disponível em <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object> Acesso em: < 25 mar. 2018 >.

TWITTER. Introduction to Tweet JSON – Twitter Developers. 2017b. Disponível em <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json> Acesso em: < 25 mar. 2018 >.

Apêndice A – Distribuição de Tweets via Etiqueta [coordinates]

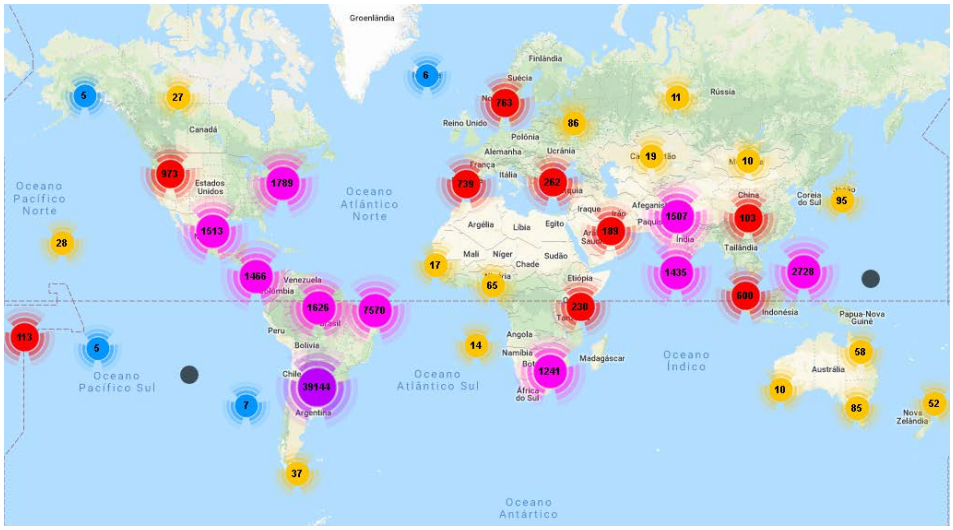


Fonte: Dados da Pesquisa, 2018.

Apêndice B – Distribuição de Tweets via Etiqueta [user]->[location]

Fonte: Dados da Pesquisa, 2018.

Apêndice C – Distribuição de Tweets via Etiqueta [place]->[bounding_box]->[coordinates]



Fonte: Dados da Pesquisa, 2018.

METAS, AÇÕES E INDICADORES COMO SUBSÍDIOS PARA ANÁLISE DA QUALIDADE DE DADOS:

Uma inversão necessária entre consequentes e antecedentes

GOALS, ACTIONS AND INDICATORS AS A CLUE TO DATA QUALITY ANALYSIS:
A necessary reversal between consequents and antecedents

Pedro Jácome de Moura Jr[1], Maicon Henrique Ferreira Aragão[(2)]
(1) Universidade Federal da Paraíba, Campus I/CCSA/DA, pjacome@sti.ufpb.br.
(2) Universidade Federal da Paraíba, Campus I/CCSA/CAdm, maiconhenriq_ilhadg@hotmail.com.

Resumo:
Assume-se que dados compõem os insumos fundamentais para a compreensão e transformação da realidade organizacional e, por conseguinte, a qualidade dos dados deve ser considerada para determinar em que medida esses dados podem ser usados para suporte a decisões. O objetivo deste estudo (em andamento) é identificar os níveis de qualidade dos dados contidos nos repositórios institucionais da Universidade Federal da Paraíba (UFPB) para suporte à mensuração dos indicadores de desempenho institucionais, realizando, para tanto, uma análise sobre o Plano de Desenvolvimento Institucional (PDI) da UFPB. Os achados demonstram que apenas 11% das metas possuem definição de indicadores para seu monitoramento, o que enseja considerações acerca da compreensão e aplicação dos conceitos de gestão e de governança por parte da instituição analisada e sugerem uma mudança na estratégia de coleta de dados para as etapas seguintes.
Palavras-chave: Governança corporativa. Governança de TI. Qualidade de dados. Planejamento estratégico.

Abstract:
This study assumes that data composes the fundamental inputs for the understanding and transformation of organizational reality and, therefore, the quality of the data should be considered to determine to what extent such data can be used to support decisions. The study (in progress) aims to identify the data quality levels exposed by the institutional repositories of the Universidade Federal da Paraíba (UFPB) to support the measurement of institutional indicators, performing an analysis of the UFPB's Institutional Development Plan (PDI). The findings show that only 11% of the defined targets have indicators for their monitoring, which leads to considerations about the understanding and application of management and governance concepts by the analyzed institution, and suggest a change in the data collection strategy for the next steps.
Keywords: Corporate governance. IT governance. Data quality. Strategic planning.

1 INTRODUÇÃO

As organizações podem ser vistas como um contínuo processo de desenvolvimento, fluxo e troca de informações entre as diversas partes que as compõem (Simon, 1950; Winter, 2003).

Se houve um período em que havia limitação de dados disponíveis para análise, hoje o volume, velocidade e variedade com que os dados são produzidos é tão maior que a dificuldade se concentra na transformação desses dados em indicadores que subsidiem a tomada de decisão e aplicação. Para alguns autores, pode-se afirmar que as organizações estão inundadas de dados (Davenport, 2006; Hazen *et al.*, 2014) e, paradoxalmente, não conseguem lidar adequadamente com isso.

Sendo o dado o insumo fundamental para a compreensão e transformação da realidade organizacional, a qualidade desse dado determina se – e em que medida – o mesmo poderá ser usado para suporte a decisões (O'Reilly, 1982), considerando também as características de veracidade (precisão) e valor (potencial de produção de benefícios) do dado (Bedeley et al., 2018).

A qualidade do dado pode ser objetivamente analisada por meio de quatro dimensões intrínsecas (independem de contexto): (1) precisão (*accuracy*), ou o grau em que o dado (registrado em um sistema de informações) corresponde à realidade de que foi obtido; (2) atualização (*timeliness*), ou frequência de obtenção de valores atuais para o dado; (3) consistência (*consistency*), ou o grau de correspondência de formato e estrutura entre dados de mesma natureza (e.g. datas); e (4) completude (*completeness*), ou a disponibilidade compreensiva de atributos em um modelo de dados para representação da realidade (Ballou *et al.*, 1998; Hazen *et al.*, 2014; Redman, 1998).

No entanto, pressupõe-se que analisar a qualidade de todos os dados disponíveis em uma organização é uma tarefa onerosa e que tal intento pode ser mal sucedido, perdendo-se entre prioridades. Considerando-se que prioridades são definidas em planejamentos estratégicos organizacionais, um início adequado para analise de qualidade de dados pode ser definido pelos dados que subsidiam a elaboração de indicadores de monitoramento da estratégia (Kaplan & Norton, 2001).

2 OBJETIVOS

O objetivo central deste estudo é identificar os níveis de qualidade dos dados contidos nos repositórios institucionais da Universidade Federal da Paraíba (UFPB) para suporte à mensuração dos indicadores de desempenho institucionais. Para isso, faz-se necessário (a) identificar os indicadores de desempenho contidos no Plano de Desenvolvimento Institucional (PDI) da UFPB, em vigência; (b) identificar os dados necessários ao monitoramento de cada indicador de desempenho; e (c) verificar a qualidade dos dados necessários à obtenção de cada indicador de desempenho.

3 PROCEDIMENTOS METODOLÓGICOS

O PDI UFPB (disponível em <http://www.proplan.ufpb.br/proplan/contents/documentos/pdi/pdi_ufpb_2014-2018.pdf>, acesso em 25/08/2018), contendo 131 páginas de texto corrido,

foi utilizado como fonte de dados secundários para esta pesquisa. Definiram-se como unidades de análise as pró-reitorias (PRG, PRPG, PRAC, Proplan, Progep, PRAPE e PRA) e os setores Agência de Inovação (INOVA), Superintendência de Tecnologia da Informação (STI) e Biblioteca Central (BC) por representarem setores de gestão, diferentemente de unidades acadêmicas (como seria o caso dos centros de ensino).

As palavras-chave "meta" e "indicador" foram aplicadas sobre o *corpus* que compõe o PDI. Os excertos contendo menção às palavras-chave foram tabulados em planilha eletrônica composta pelas colunas localização (página/linha), setor, eixos/objetivos, metas, indicadores e ações. A análise do termo "ação" ou "ações" se mostrou necessária, embora não prevista inicialmente, em função de aparentes sobreposições conceituais com o termo "meta".

Cada excerto foi analisado e codificado como (categorias definidas *a priori*) "meta" ou "indicador"; e como (categoria emergente) "meta definida como ação".

Para fins metodológicos, adota-se como definição de meta o estabelecimento de objetivo específico (sem ambiguidades), mensurável (medida de alcance da meta), apropriado (relevante para a estratégia), realista (contextualizada em relação a restrições e desempenho conhecido) e com prazo determinado (conceito de metas "SMART"; Rietbergen & Block, 2010).

Assume-se como definição de indicador de desempenho "um número, percentagem ou razão que mede um aspecto do desempenho, com o objetivo de comparar esta medida com metas preestabelecidas" (da União, 2010). Ou seja, um indicador de desempenho compara o desempenho real com algum parâmetro (dado previamente conhecido ou organização da mesma natureza).

4 RESULTADOS E DISCUSSÃO

O Apêndice "A" sintetiza quantitativamente os achados. Observa-se que dentre os dez setores analisados foram identificados 132 metas e apenas 15 indicadores a estas relacionados.

As metas identificadas demandaram distinção em relação ao uso do termo "ação" (ou "ações"), pois durante a análise verificou-se que houve sobreposição quanto às definições (por exemplo, nas págs. 36 e 37 as ações da Proplan são descritas como metas).

A exposição das metas no PDI também encontra formato distinto dentre os setores, já que alguns optaram por destaca-las em tópicos, o que possibilitou melhor clareza para sua identificação, enquanto outros optaram por mantê-las sem destaque, junto ao texto corrido (por exemplo, metas descritas pela PRPG na forma "A expansão quantitativa se dará a partir de estudos prospectivos e de diagnóstico da demanda", pág. 20). Provavelmente essa multiplicidade de forma se deve à estratégia de desenvolvimento do PDI, em que cada setor elaborou individualmente a sua parte, resultando por fim uma "compilação" vagamente integrada.

Verifica-se que o número de indicadores identificados (15) não corresponde ao número de metas identificadas (132). Uma tal discrepância sugere múltiplas interpretações, dentre as quais pode-se destacar, principalmente, sobreposição (intencional ou não, pois não há meios, nesta pesquisa, para afirmação inequívoca) dos conceitos de gestão e de governança. Ora, se uma meta é estabelecida por uma gestão e não há meios de verificação de sua execução, por suposto não há meios de exercício da governança, de modo que qualquer evidência de alcance da meta passa a ser questão de fé.

Por governança entende-se o conjunto de políticas e práticas que explicitem a atuação da gestão, sendo a transparência de ações um dos principais mecanismos de governança. No entanto, ao sobrepor os conceitos de gestão e governança, este último perde sua conotação (torna-se abstrato e inefetivo), "fazendo com que organizações pratiquem governança por imposição ou imitação, [apenas] como uma forma de 'dar satisfação' aos órgãos de controle" (De Moura Jr., 2017, p. 3).

Não obstante, destaca-se que os setores Proplan, PRPG e STI estabeleceram como meta a criação ou melhoria dos indicadores de desempenho, como se pode observar em excertos do tipo "ampliar e modernizar laboratórios de pesquisa, equipamentos de informática, aquisição de softwares, material de consumo, oferecendo condições para melhoria dos indicadores de desempenho dos programas" (PRG; PDI, 2014; p. 21), "Definição de indicadores de desempenho da atividade de TI na instituição" (STI; PDI, 2014; p. 41) ou "Construir e acompanhar os indicadores de desempenho acadêmico" (Proplan; PDI, 2014; p. 37).

Observou-se que muitas definições de metas são falhas por não conter a ação específica que se deve empreender, vide "Estabelecimento da Formação continuada de docentes" (Progep; PDI, 2014; p. 37) ou "implantação dos módulos do Sistema Integrado de Gestão de Recursos Humanos – SIGRH" (STI/Progep; PDI, 2014; p. 39). Ações definidas como verbos no infinitivo (realizar, implementar, elaborar, por exemplo), por outro lado, fornecem uma indicação direta do que deve ser feito, sem ambiguidades, conforme Rietbergen & Block (2010).

Em relação ao alcance do objetivo central deste estudo, qual seja, identificar os níveis de qualidade dos dados contidos nos repositórios institucionais da UFPB para suporte à mensuração dos indicadores de desempenho institucionais, tem-se que as dificuldades enfrentadas com o primeiro desdobramento específico (identificar os indicadores de desempenho contidos no PDI da UFPB, em vigência) ainda não permitem avançar consistentemente sobre os passos seguintes.

Identificar os dados necessários ao monitoramento de apenas 15 indicadores de desempenho não parece justificável. Assim, estima-se que a investigação sobre os indicadores deva avançar em consulta aos gestores que definiram as metas, ou seja, aqueles gestores que participaram da elaboração do PDI ou aqueles que vêem conduzindo seus setores seguindo a orientação contida no PDI.

Pretende-se com essas consultas, entrevistar os gestores apresentando-lhes os resultados parciais aqui descritos e obter deles definições pormenorizadas da relação entre metas e indicadores e de indicadores e dados.

Somente a partir de uma definição mínima de indicadores, poder-se-á avançar para as análises de qualidade dos dados que os subsidiam.

4 CONSIDERAÇÕES FINAIS

Este estudo em andamento investigou, até o momento, o Plano de Desenvolvimento Institucional (PDI) da UFPB (que aqui serve como uma ilustração de documento estratégico) em busca de definições de metas e indicadores de desempenho associados.

Os achados demonstram que apenas 11% (15/132) das metas possuem definição de indicadores para seu monitoramento. Este baixo índice enseja considerações acerca da compreensão e aplicação dos conceitos de gestão e de governança por parte dos elaboradores do PDI da instituição

analisada, além de provocar curiosidade quanto à capacidade de demonstração de cumprimento de metas por parte da gestão que adota um tal documento estratégico.

Ainda quanto aos achados, embora esta pesquisa enderece uma perspectiva específica de qualidade dos dados, os documentos que servem de fonte demandam atenção sobre aspectos que antecedem essa perspectiva específica, a saber: compreensão, definição e uso adequado dos termos "meta", "ação" e "indicador".

Os próximos passos envolvem entrevistas com gestores que participaram da elaboração do PDI e com aqueles que vêem conduzindo seus setores seguindo as orientações contidas no PDI. Em seguida, compreendendo melhor os indicadores em uso (já que aqueles formalmente declarados são insuficientes em seu estado atual), poder-se-á identificar os dados necessários à elaboração de cada indicador e, por fim, analisar a qualidade desses dados.

É nesse sentido que este artigo pratica uma inversão na ordem tradicional de investigação da temática. Enquanto tradicionalmente tem-se dados (com suposta qualidade) que dão suporte a indicadores de desempenho que, por sua vez, permitem o monitoramento do alcance de metas, aqui, de acordo com as circunstâncias, observam-se metas que devem (ou deveriam) ter indicadores associados, formados por dados cuja qualidade demanda análise.

REFERÊNCIAS

Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science*, *44*(4), 462-484.

Bedeley, R. T., Ghoshal, T., Iyer, L. S., & Bhadury, J. (2018). Business analytics and organizational value chains: a relational mapping. *Journal of Computer Information Systems*, *58*(2), 151-161.

Brasil. UFPB. *Plano de Desenvolvimento Institucional* (PDI) 2014-2018. Disponi□vel em: <http://www.proplan.ufpb.br/proplan/contents/documentos/pdi/pdi_ufpb_2014-2018.pdf>. Acesso em: 28/08/2018.

da União, T. D. C. (2010). Técnica de indicadores de desempenho para auditorias. Portaria TCU. Disponível em: <http://portal3. tcu. gov. br/portal/page/portal/TCU/comunidades/programas_governo/te cnicas_anop/BTCU_indicadores_de_desempenho.pdf>. Acesso em 28/08/2018.

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard Business Review*, *90*(5), 70-76.

De Moura Jr., P. J. (2017). Governança de Tecnologia da Informação: A meio caminho entre o isomorfismo e a comoditização. *Revista Eletrônica de Sistemas de Informação*, *16*(3), no prelo.

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, *154*, 72-80.

Kaplan, R. S., & Norton, D. P. (2001). Transforming the balanced scorecard from performance measurement to strategic management: Part I. *Accounting horizons*, *15*(1), 87-104.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, *34*(3), 387-394.

O'Reilly, C. A. (1982). Variations in decision makers' use of information sources: The impact of quality and accessibility of information. *Academy of Management Journal*, *25*(4), 756-771.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79-82.

Roberts, N., Galluch, P. S., Dinger, M., & Grover, V. (2012). Absorptive capacity and information systems research: Review, synthesis, and directions for future research. *MIS quarterly*, *36*(2), 625-648.

Rietbergen, M. G., & Blok, K. (2010). Setting SMART targets for industrial energy use and industrial energy efficiency. *Energy Policy, 38*(8), 4339-4354.

Simon, H. A. (1950). *Administrative behavior*. New York: Macmillan.

Winter, S. G. (2003). Understanding dynamic capabilities. *Strategic management journal*, *24*(10), 991-995.

Apêndice A – Quantitativo de metas, indicadores e ações identificados

Setores	Metas	Indicadores de Desempenho	Metas descritas em ações
PRAC	28	4	29
PRG	27	3	-
Progep	24	-	11
PRPG	11	3	4
INOVA	10	2	-
STI	10	-	-
PRAPE	9	3	9
BC	8	-	-
Proplan	5	-	5
PRA	-	-	-
Total	132	15	58

Fonte: Dados da pesquisa (2018)

MÉTODO MULTICRITÉRIOS PARA AUXILIAR O PROCESSO DE TOMADA DE DECISÃO SOBRE EMPREENDIMENTOS IMOBILIÁRIOS UTILIZANDO MODELOS BIM

MULTICRITERIA METHOD TO ASSIST THE DECISION-MAKING PROCESS ON REAL ESTATE UNDERTAKINGS USING BIM MODELS

Bruno Cesarino Soares[1], Rogério Amaral Bonatti[2], Renata Maria Abrantes Baracho[3]
(1) Universidade Federal de Minas Gerais, brunocesarino@hotmail.com
(2) Universidade Federal de Minas Gerais, rbonatti@gmail.com
(3) Universidade Federal de Minas Gerais, renatabaracho@eci.ufmg.br

Resumo:
O presente trabalho apresenta a utilização de informações obtidas em modelos de informações da construção (*Building information Modeling* – BIM) e a avaliação de empreendimentos imobiliários, em uma estrutura multicritérios. As informações sobre os possíveis projetos de construção, reforma, ampliação, aquisição, ou aluguel de edificações são utilizadas para viabilizar um processo de tomada de decisão. Os modelos funcionam como uma base para a representação digital de dados ou informações de uma construção e trazem diversas possibilidades de análise. Assim, a análise multicritério classifica os empreendimentos imobiliários representados como modelos BIM. A hierarquização do problema envolve as relações entre metas, critérios que expressam metas e alternativas. A análise comparativa entre os critérios considerados e a interação com os elementos que compõem os modelos de projetos de construção proporcionou a execução prática de um problema que interliga dados de ambientes distintos com um método de análise multicritérios. A avaliação de viabilidade do processo construtivo reforça a necessidade de expansão da metodologia e da amostra, para se definir os elementos que compõem a tomada de decisão no setor de empreendimentos imobiliários.
Palavras-chave: Métodos Multicritérios. Analytic Hierarchy Process (AHP). Modelagem de Informações da Construção. Building Information Modeling (BIM).

Abstract:
This work presents the development of a multicriteria method to evaluate real estate projects based on information retrieved from Building Information Modeling (BIM). Information on possible construction, renovation, expansion, acquisition or rental of buildings is used to enable a decision-making process. The models work as a basis for the digital representation of data or information of a construction and bring diverse possibilities of analysis. Thus, the proposed method classifies the real estate ventures represented as BIM models. Complex problems in a hierarchical structure involve relationships between goals, criteria that express goals, and alternatives. Given this complexity, the proposed analysis method evaluates hierarchical structures expressed from objectives, or criteria, and alternatives. Thus, the Analytic Hierarchy Process (AHP) is used, which occurs at successive levels, beginning with the construction of hierarchies, followed by the definition of priorities and finally a verification of the logical consistency in order to arrive at a comparative result between the alternatives.
Keywords: Multicriteria Methods. Analytic Hierarchy Process (AHP). Building Information Modeling (BIM).

1 INTRODUÇÃO

Com a constante demanda por utilização de novos espaços e tendo em vista a disponibilidade de ativos imobiliários, o gerenciamento de informações pode ser muito útil no processo de tomada de decisão na área de Arquitetura Engenharia e Construção (AEC).

Novos projetos imobiliários envolvem diversas atividades como aquisição, construção, ampliação, reforma ou aluguel de edificações. Assim, a decisão acerca de qual empreendimento realizar deve ser pautada na elaboração de projetos com os adventos tecnológicos utilizados desde sua concepção, podendo ser beneficiada com a melhoria da Gestão de Projetos em todas suas fases.

Os atuais recursos de Tecnologia da Informação e Comunicação (TIC) permitem que os processos de elaboração de projeto ou desenho assistido por computador (*Computer-Aided Design* – CAD), incorporem diversos recursos informacionais, indo muito além de um simples uso de computadores como ferramentas gráficas para a criação de desenhos técnicos.

Nesse contexto, além de fornecer ferramentas para a representação de formas geométricas utilizando como base a tecnologia CAD, o processo conhecido como modelagem de informações da construção (*Building Information Modeling* – BIM) procura otimizar a elaboração de projetos de construções.

Para Chiavenato (2004), o planejamento possibilita a determinação antecipada das atividades que devem ser desempenhadas, dos objetivos serão alcançados e auxilia as organizações a se organizarem para atingirem as metas pretendidas.

Novas ferramentas que se baseiam em desafios de projetos em geral, tais como orçamentos, prazos e excelência do produto, podem auxiliar as organizações a obterem melhores resultados e à aproximação do objeto executado ao proposto no planejamento inicial.

De acordo com o Instituto de Gerenciamento de Projetos *(Project Management Institute* – PMI[1]), "o Gerenciamento de Projetos é a aplicação de conhecimentos, habilidades e técnicas para a execução de projetos de forma efetiva e eficaz." As informações dos modelos utilizados, preenchidas durante todo o ciclo de vida dos projetos, fornece e define um ciclo de vida dos dados, o que possibilita a avaliação de importâncias em momentos distintos.

A identificação de informações de várias fontes externas e internas a projetos da área de AEC possibilita a inclusão de dados nos modelos digitais gerados para representar esses projetos (modelos BIM). Destes modelos foram considerados três para a estruturação dos problemas de forma hierárquica. A análise dos critérios, então, demonstra uma dentre as possibilidades de se chegar a uma solução desejada, baseada em comparações par-a-par entre eles, de um processo que por vezes se torna imprescindível para a tomada de decisão.

2 OBJETIVOS

O objetivo deste trabalho é viabilizar a utilização da modelagem de informações da construção (BIM) em um processo de tomada de decisão em empreendimentos imobiliários por meio de uma análise multicritérios. O método de análise a ser adotado utiliza a identificação dos

1 Project Management Institute – PMI. Disponível em: <https://brasil.pmi.org/>. Acesso em: 10 set. 2018.

elementos que compõem a estrutura de decisão e pretende realizar uma comparação das possíveis alternativas de projetos da área. Considera-se os dados sobre empreendimentos dos modelos BIM como premissa para a automação do preenchimento de alguns critérios de análise exigidos pelo método. É esperado, ainda, avaliar os resultados obtidos com relação ao processo decisório e verificar a viabilidade de continuidade da pesquisa, com melhorias e ajustes no método, bem como sua aplicação a outros cenários.

3 PROCEDIMENTOS METODOLÓGICOS

3.1 Métodos Multicritérios

A modelagem de problemas complexos em uma estrutura hierárquica envolve as relações entre metas, critérios que expressam metas e alternativas. A estrutura hierárquica expressa a partir do objetivo ou objetivo para os critérios e alternativas em níveis sucessivos (SAATY & VARGAS, 2001).

O método utilizado possui três princípios para sua aplicação. A construção de hierarquias, em que o problema é estruturado em níveis hierárquicos e é um passo fundamental para compreendê-lo. A definição de prioridades, que se baseia na capacidade de perceber as relações entre objetos e situações diversas. A comparação entre os pares de critérios e a consistência lógica, em que o *Analytic Hierarchy Process* (AHP) permite avaliar o modelo de priorização construído quanto à sua consistência, fazendo uso de conceitos matemáticos, principalmente princípios e propriedades de matrizes.

Para este trabalho, optou-se por um método de análise multicritério que fosse do domínio dos autores, possibilitando a execução e para a avaliação da viabilidade da pesquisa. Uma vez considerada executável e viável, os dados provenientes das entrevistas não estruturadas vão ser relacionados entre si e entre as informações ou os critérios advindos dos modelos BIM. Isso permite a construção e a triangulação das evidências traduzidas pelo estudo de caso, apresentado em trabalhos anteriores (BONATTI; BARACHO; 2015, 2016). A descrição e as etapas de cálculos estão descritas nestes trabalhos. A definição de prioridades entre os critérios é realizada entre os elementos da hierarquia identificada. São geradas as matrizes de julgamento (sempre matrizes quadradas), onde o número na linha ***i*** e na coluna ***j*** dá a importância do critério $\mathbf{C}_{i}$ em relação à $\mathbf{C}_{j}$. A importância relativa entre os elementos visa minimizar a inconsistência da elaboração das matrizes de julgamento. O grau de inconsistência reflete a adequação dos julgamentos realizados pelos gestores e um grau máximo é estipulado, também descrito nos trabalhos anteriores.

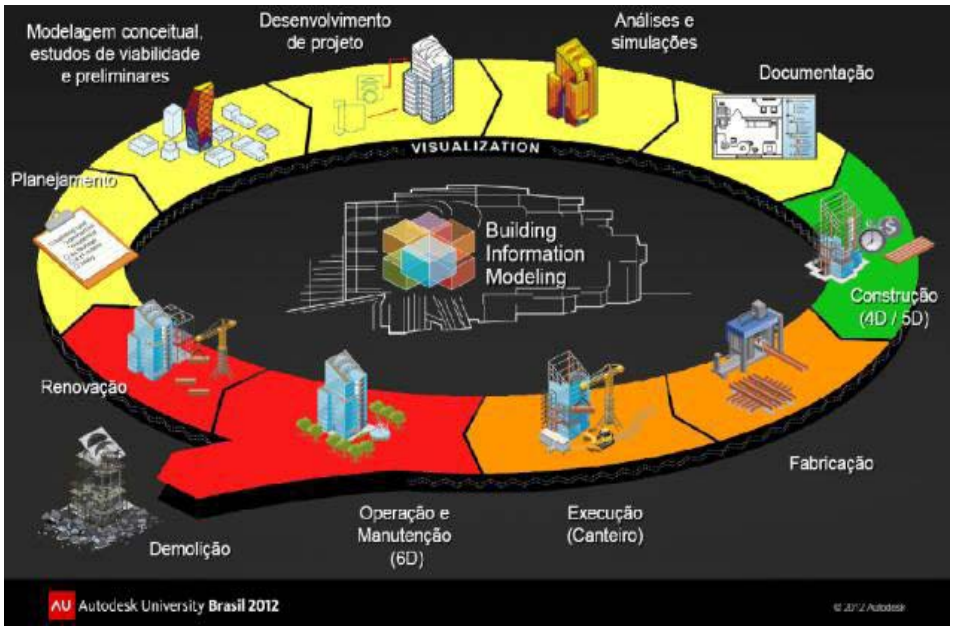
3.2 RECUPERAÇÃO DA INFORMAÇÃO EM AMBIENTE BIM

O processo tecnológico conhecido como modelagem de informações da construção (BIM) surgiu com a evolução das aplicações gráficas voltadas para a geração de desenhos técnicos

(CAD) ocorrida na década de 1990. Essa nova forma de se projetar, ou melhor, de modelar as informações de uma construção, bem a aplicação de soluções CAD à indústria de AEC são amplamente discutidas em trabalhos como Eastman *et al.* (2005) e Björk e Laakso (2010). Algumas das principais características, recomendações e formas de uso das diversas tecnologias associadas ao BIM, ou simplesmente tecnologia BIM, podem ser obtidas em Eastman *et al.* (2011) e em Porto *et al.* (2015), em que é apresentado um detalhamento das características e recomendações para a implantação e melhor aproveitamento dessa tecnologia. O BIM também é amplamente discutido em trabalhos como os de Jacoski (2003 e 2008), Flemming *et al.* (2004), Isikdag e Underwood (2009), Pereira Junior e Baracho (2015) e Laiserin (2018), e pode ser considerado um novo padrão da indústria, assim como aconteceu com a tecnologia CAD.

Com o uso da tecnologia BIM, todas as informações necessárias em um projeto de construção ficam contidas em um modelo digital ou virtual, e ela pode ainda ser utilizada na automação da apresentação de documentos, melhorando aspectos como a construtibilidade, a análise de conflitos, o planejamento, a elaboração de cronogramas (4D), a análise de custos e orçamentação (5D), entre outros. Vale ressaltar que o ciclo de vida do processo de BIM abrange toda a vida útil de uma edificação ou construção, conforme mostrado na Figura 1, retirada de Mello (2012).

Figura 1 – Processo BIM (Mello, 2012)



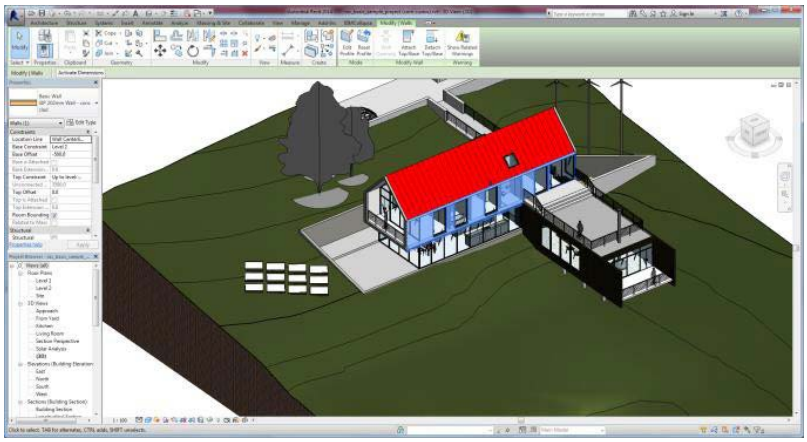
O BIM pode ser entendido como a tecnologia responsável pela gestão de uma enorme quantidade de informações relativas às construções, sendo também um grande repositório de informações. Nesse contexto, o uso da tecnologia BIM no mundo traz desafios e oportunidades para a realidade brasileira (WATSON, 2011; BRYDE *et al.*, 2013). Essa tecnologia também já está se consagrando e ganhando destaque como ferramenta para elaboração e gestão de projetos de Arquitetura e Engenharia no Brasil (PORTO *et al.*, 2015; PEREIRA JUNIOR; BARACHO, 2015).

Os *softwares* BIM são executados em plataformas ou ambientes de CG e utilizam diversos recursos de TIC para permitir a manipulação de modelos digitais de uma edificação ou construção. Esses *softwares* utilizam as representações digitais multidimensionais e paramétricas,

bem como, intrinsecamente, os paradigmas da programação orientada a objetos (POO), que é uma tecnologia para programação de computadores em que "objetos" são os elementos fundamentais que representam as entidades envolvidas no sistema. Mais detalhes sobre a POO podem ser obtidos em trabalhos como o de Gamma *et al.* (2000).

A Figura 2 apresenta um modelo BIM criado com o uso do *software* Revit.

Figura 2 – Modelo BIM com Componente Construtivo ("família de parede") Selecionado

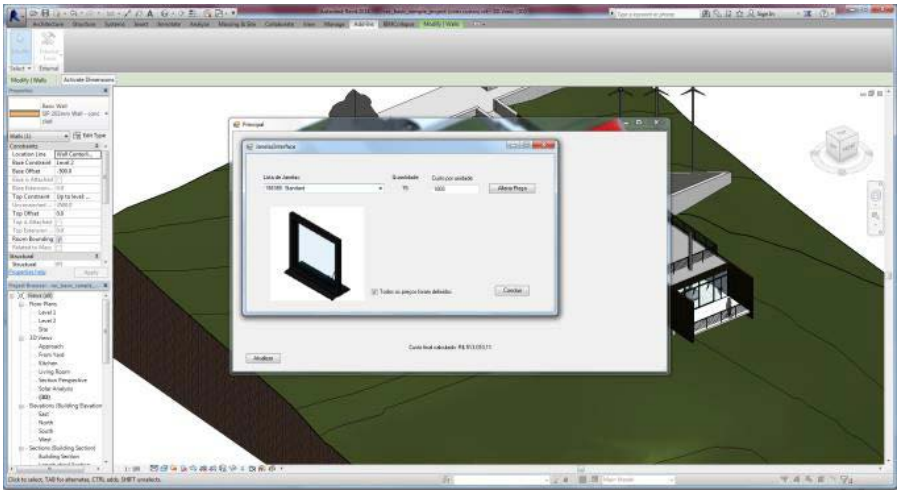


Muitos dos *softwares* que utilizam a tecnologia BIM também oferecem ferramentas para se estender suas funcionalidades, mediante a criação de novas interfaces dentro de seus ambientes (aplicativos de extensão conhecidos como *plug-ins* ou *add-ins*), aumentando ainda as capacidades do BIM.

Esses aplicativos de extensão podem ser desenvolvidos com vistas a acessar, via código-fonte, os componentes paramétricos dos modelos BIM, que representam as partes construtivas, e com isso recuperar informações úteis em diversos contextos. Assim, as informações contidas nos modelos BIM permitem diversas análises, como a proposta neste trabalho.

A Figura 3 mostra um protótipo desenvolvido para ler dados de um modelo BIM e preparar um orçamento global da obra com base nos custos de cada componente construtivo.

Figura 3 – Aplicativo de Extensão Criado para Calcular o Orçamento de um Modelo BIM



Nota-se que a elaboração de soluções computacionais para problemas de Engenharia utilizando a tecnologia BIM pode se associar ao uso de diversas ferramentas e metodologias. Alguns trechos de códigos-fonte, úteis para o desenvolvimento de aplicativos de extensão para o *software* Revit, se encontram disponibilizados em páginas da *internet*, como Tammik (2018).

Tendo em vista esses exemplos, pretende-se desenvolver futuramente um aplicativo de extensão para recuperar automaticamente informações dos modelos BIM e alimentar o os critérios a serem utilizados pelo *software* que implementa a análise multicritérios proposta. Nessa análise, cada modelo BIM, que se refere a um empreendimento, é então avaliado de acordo com uma comparação feita por especialistas da área.

4 RESULTADOS

As informações contidas no contexto de um processo e de modelos BIM, recuperadas a partir de funcionalidades presentes em uma ferramenta ou *software*, foram insumos para a execução e o resultado apresentado. Com o uso destas informações, o método multicritérios permitiu a análise comparativa entre os critérios considerados e foram considerados três cenários de projetos. Também foi realizada uma avaliação dos resultados obtidos.

4.1 Análise Multicritério

Os critérios foram definidos a partir das informações extraídas dos modelos BIM acrescidos de outros, baseados em considerações de respondentes especialistas do setor de AEC. A avaliação desses critérios e o cálculo de consistência permitem organizar os resultados de cada um dos respondentes. Para a execução, as matrizes de julgamento são processadas, e tal processo define a prioridade de um empreendimento em detrimento aos demais.

4.2 Definição dos Critérios

Em um ambiente BIM, alguns critérios podem ser obtidos automaticamente a partir dos modelos. Além, foram considerados critérios subjetivos, por exemplo, a intenção do gestor de empreender projetos em determinada área ou que utilize certos materiais ou mesmo que ele julgue ter mais expertise para executar. Assim, foram propostos os seguintes critérios: **C1: Custo global da obra ou do empreendimento**: critério objetivo, obtido com base no modelo BIM; **C2: Nível de detalhamento do projeto/modelo**: este critério utiliza uma escala própria dos modelos, mas normas BIM podem ser aplicadas; **C3: Complexidade do modelo**: considera-se algumas variáveis internas do modelo BIM, como tipo e quantidade de materiais e existência de componentes com geometria curva. Outras variáveis externas, como tipo de terreno, tipo de fundação e agressividade do ambiente também poderiam ser adotadas; **C4: Maturidade da**

organização com relação ao uso do BIM: não possui um valor especificado, pois avalia aspectos como a experiência da organização empreendedora, a quantidade de profissionais treinados e a existência de *software* apropriado; **C5: Experiência do empreendedor**: critério que considera o interesse pessoal/organizacional pelo empreendimento.

4.3 Avaliação

Os dados criados para avaliar o modelo proposto são dados computacionais, uma vez que a entrada de dados foi definida a partir de modelos criados para testes e os resultados foram obtidos a partir de simulações utilizando o *software* Revit, paraos modelos BIM. Os três empreendimentos a serem avaliados são **A1: projeto de uma casa com dois pavimentos a ser reformada**. O nível de detalhamento do modelo BIM é muito baixo. Materiais padrão foram adotados e a geometria modelada é apenas aproximada, mas o custo é bem baixo; **A2: projeto de uma nova edificação com cinco pavimentos**. O nível de detalhamento do modelo BIM é mediano e a especificação é de materiais de referência, mas as geometrias refletem a realidade, devendo, portanto, haver poucas mudanças durante a execução. Entretanto não foi feita a atribuição de custos, e existe um limite de orçamento; **A3: projeto de nova casa com dois pavimentos**. O modelo BIM possui nível de detalhamento bastante alto, com cobertura total de componentes construtivos. A especificação de materiais reflete exatamente o que deve ser executado e foi feita a atribuição de custos a todos componentes, porém, o custo executivo foi o mais alto entre os projetos.

A partir deste cenário, um especialista participou da simulação. Com escala de conformidade de valores (1 - igual, 2 - pouco mais importante, 3 - mais importante e 5 - muito mais importante), chegou-se à matriz:

Tabela 1 – Correlação entre Critérios

C1	C2	C3	C4	C5
1	2	3	1/3	1/2
1/2	1	3	1/2	1
1/3	1/3	1	2	3
3	2	1/2	1	1
2	1	1/3	1	1

A matriz reflete as comparações de pares. Para exemplificar, em um objetivo global (novo empreendimento), o custo (C1) é quanto importante em relação ao detalhamento do modelo BIM (C2)? Assim, as importâncias (ponderação) dos critérios são determinadas em relação ao objetivo. Denomina-se a Prioridade Global a normalização desta matriz e, seguindo os passos metodológicos, define o vetor de prioridade global (VPG).

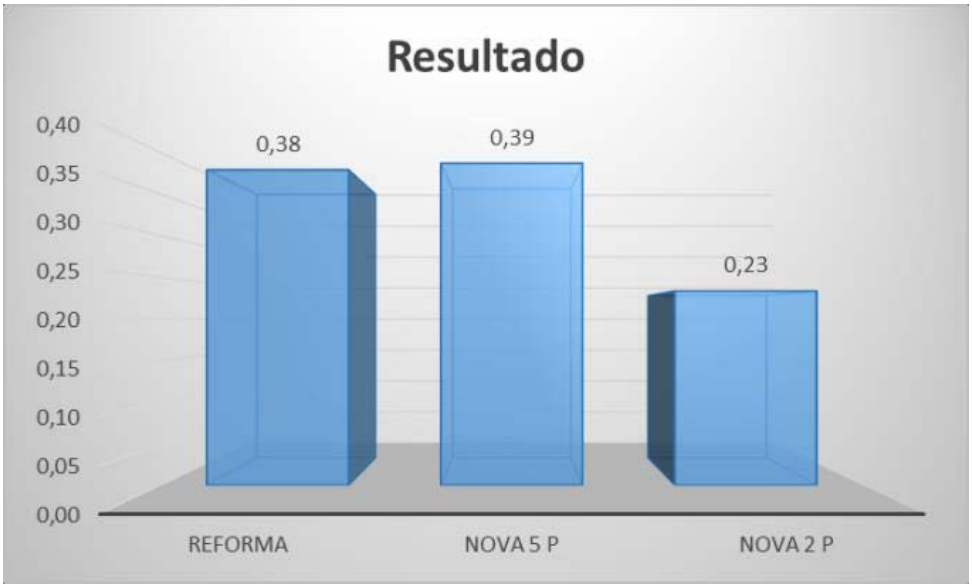
Definem-se então as importâncias dos critérios em relação às alternativas. A Prioridade Média Local (PML), gerando sua matriz correspondente. A normalização e o procedimento metodológico definem os vetores de PMLs:

Tabela 2 – Exemplo de Correlação entre Alternativas de Empreendimentos (C1)

A1	A2	A3
1	2	5
1/2	1	3
1/5	1/3	1

O somatório da multiplicação das PMLs pelo VPG é a última etapa do cálculo definido pela análise e gera o resultado que define a Prioridade Global, como se segue.

Figura 4 – Resultados da Análise



A prioridade global reflete a importância dada aos critérios envolvidos em relação ao objetivo global e em relação às alternativas, expressando em números as considerações dos respondentes.

5 CONSIDERAÇÕES FINAIS

No contexto informacional, o BIM funciona como uma base para a representação digital de dados ou informações de uma construção, não sendo apenas um desenho ou projeto estático. Os modelos representam os componentes dos sistemas construtivos de forma dinâmica, podendo se interligar a diversos outros repositórios ou sistemas de informação.

O método proposto apresenta uma forma de se explorar o potencial do BIM em conjunto com uma avaliação subjetiva, adotando critérios intangíveis. As análises realizadas verificaram o funcionamento do método, mas ressalte-se que foram utilizados dados computacionais, gerados ou obtidos com base em modelos criados especificamente para testes. Com isso sugere-se como trabalho futuro que os dados de entrada dos modelos sejam adaptados à realidade de possíveis utilizadores desse método, que devem avaliar os resultados e calibrar as entradas de acordo com as particularidades de suas organizações.

REFERÊNCIAS

CHIAVENATO, Idalberto. **Administração nos novos tempos.** 2. ed. Rio de Janeiro: Elsevier, 2004.

BONATTI, R. A.; BARACHO, R.M.A.; BARACHO, F. R. A. C.; PESSANHA, C. P.; REZENDE, M. M. S.; LIMA, F. B.; SILVA, C. H. F. **Análise de projeção e viabilidade técnica de novos empreendimentos para geração de energia elétrica.** Anais... Enancib 2016, 2016.

BONATTI, R. A., BARACHO, R. M. A. **A gestão da informação e o processo decisório no setor energético:** mensuração de critérios e alternativas. *Pesquisa Brasileira em Ciência da Informação e Biblioteconomia*, v. 10, p. 237-249, n. 2015.

BJÖRK, B-C., LAAKSO, M., **CAD standardisation in the construction industry** – A process view. Automation in Construction. v.19, n.4, 398-406 pp, 2010.

BRYDE, D., BROQUETAS, M., VOLM, J.M., **The project benefits of Building Information Modelling (BIM)**, International Journal of Project Management. Volume 31, Issue 7, Pages 971-980, 2013. ISSN 0263-7863.

EASTMAN, C. M., WANG, F., You, S. -J., Yang, D., **Deployment of an AEC industry sector product model.** Computer-Aided Design. v.37, n.12, 1214-1228 pp, 2005.

EASTMAN, C. M., TEICHOLZ, P., SACKS, R., LISTON, K., **BIM Handbook:** A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers, and Contractors – Second Edition. John Wiley & Sons, New Jersey. 2011.

FLEMMING, U., ERHAN, H., ÖZKAYA, I., **Object oriented application development in CAD:** a graduate course. Automation in Construction. v.13, 147-158 pp, 2004.

GAMMA, E., HELM, R., JOHNSON, R., VLISSIDES, J., **Padrões de Projeto: Soluções Reutilizáveis de Software Orientado a Objetos**. Porto Alegre, RS: Bookman. 363p. Trad.: Salgado, L. A. M., 2000.

ISIKDAG, U., UNDERWOOD, J., **Two design patterns for facilitating Building Information Model-based synchronous collaboration**, Automation in construction, 2009.

JACOSKI, C. A., **Integração e Interoperabilidade em Projetos de Edificações** – uma implementação com IFC/XML. 2003. 217f. Tese (Doutorado em Engenharia de Produção e Sistemas) - Universidade Federal de Santa Catarina. Florianópolis, SC, Brasil, 2003.

_____, **A Integração da Comunicação em Projetos de Construção Utilizando Tecnologias da Informação.** In: Conferência Construção, Portugal. FEUP- Porto/Portugal, 2004.

_____, **O Intercâmbio de Dados entre SIG e Projetos de Edificações** – A Busca pela Interoperabilidade, Universidade Comunitária Regional de Chapecó. 2008. Disponível em <http://claudio.jacoski.googlepages.com/artigo_jacoski_sig ac.pdf>. Acesso em 23 mar. 2017.

LAISERIN, J., **The BIM Page**, The Laiserin Letter. 2018. Disponível em <http://www.laiserin.com>. Acesso em 29 set. 2018.

MELLO, R., **BIM e Custos:** maximize os dados do modelo com o Navisworks e o Quantity Takeoff. Autodesk University Brasil, 2012. Disponível em: <http://static-wd.autodesk.net/content/dam/au/Brasil2014/documents/materialapoio/2012/AUBR-44_Apostila.pdf>. Acesso em: 29 set. 2018.

PEREIRA JUNIOR. M. L., BARACHO, R. M., **Relações entre a Gestão da Informação e do Conhecimento e Uso de Sistema BIM por Arquitetos e Engenheiros.** 4ª Seminário Ibero-americano Arquitetura e Documentação. Belo Horizonte, 2015.

PORTO, M. F., FRANCO, J. R. Q., BARACHO, R. M. A., **Paradigma de Utilização da Tecnologia BIM para Projeto Arquitetônico e de Engenharia.** 4ª Seminário Ibero-americano Arquitetura e Documentação. Belo Horizonte, MG, Brasil. 2015.

SAATY, T. L.; VARGAS, L. G. **Models, Methods, Concepts & Applications of The Analytic Hierarchy Process.** Boston: Kluwer Academic Publishers. 2001.

TAMMIK J., **The Building Coder.** Blogging about the Revit API. 2018. Disponível em: <http://thebuildingcoder.typepad.com>. Acesso em 29 set. 2018.

WATSON, A., **Digital buildings** – Challenges and opportunities, Advanced Engineering Informatics, Volume 25, Issue 4, Pages 573-581, 2011. ISSN 1474-0346.

NARRATIVAS DE VISUALIZAÇÕES GUIADAS POR DADOS: *BIG DATA* E INOVAÇÃO

DATA-DRIVEN VISUALIZATIONS NARRATIVES: Big Data and Innovation

Adriana Alves Rodrigues[1]
(1) Universidade Federal da Paraíba, João Pessoa, adrianacontemporanea@gmail.com

Resumo:
Analisa as narrativas de visualização de dados do prêmio *Data Journalism Awards 2018*, bem como seus processos de inovação a partir do cenário de *Big Data*. A metodologia, de caráter exploratório, compreende um Estudo de Caso e discussão teórico-conceituais e empíricas dessas narrativas e suas dimensões no uso de dados a partir de revisão de literatura e aplicação de Ficha de Observação na coleta de dados e análise quali-quantitativa. Os resultados, que fazem parte de uma pesquisa maior em andamento na tese doutoral, demonstram que as narrativas guiadas por dados se constituem em uma forma diferenciada de contar histórias no jornalismo em interface interdisciplinar com a Computação, Design e Ciência da Informação. Conclui-se que a emergência do *Big Data* e da Visualização de Dados aponta para novo formato de narrativa em que os dados exercem função central.
Palavras-chave: Visualização de dados. Narrativas. Dados. Big Data.

Abstract:
This paper presents an analysis of data visualization narratives from the 2018 Data Journalism Awards, as well as their innovation processes from the Big Data scenario. The exploratory methodology comprises a case-study and theoretical-conceptual and empirical discussion of these narratives and their dimensions in the use of data from a literature review and the application of an observation sheet in the data collection and qualitative-quantitative analysis. The results, which are part of a larger research project underway in the doctoral thesis, show that data-driven narratives constitute a new way of telling stories in journalism in an interdisciplinary interface with Computing, Design and Information Science. The research finds out that the emergence of Big Data and Data Visualization points to a new narrative format in which the data play a central role.
Keywords: Data Visualization. Narratives. Data. Big Data.

1 INTRODUÇÃO

Em um contexto permeado por grandes volumes de dados, identificados pelo conceito de *Big Data*, as produções de visualizações de dados incorporaram novas linguagens e formatos, tendo como desdobramento a emergência de um novo fenômeno de natureza interdisciplinar. Para Lagoze (2014), o fenômeno do *Big Data* não se restringe a grandes volumes de dados, mas sim, a resultado da mescla de dimensões e características diferenciadas cujos efeitos podem ser revolucionários na produção de conhecimento. Neste sentido, a visualização de dados é uma área que cada vez mais ganha espaço nos debates profissionais e acadêmicos nos mais distintos campos do saber. Para além dos fatores tecnológicos que estão presentes, muitas das visualizações de dados se pautavam, até então, a apenas "mostrar os dados" (TUFTE, 2001), ou seja, eram produções que se limitavam em explorar aqueles dados por si só. No entanto, a agregação de novas tecnologias e estratégias de estruturação de dados e, principalmente, do *Big Data*, alguns trabalhos vêm apontando como aspecto inovador o fato desses dados conterem elementos de uma narrativa (SEGEL; HEER, 2010; CAIRO, 2011; VIÉGAS, 2013; KOSARA; MACKINLAY, 2013; FIGUEIRAS, 2013) e níveis de complexidade (com os cruzamentos de dados e volume) e deste modo se apresentarem de uma forma mais dinâmica. Viégas (2013) afirma que as visualizações têm a capacidade de captura de um determinado fato da realidade com profundidade através dos aspectos gráficos e diferentes cruzamentos de dados. Essa exploração do formato de visualização de dados tem ocorrido no campo científico e em organizações jornalísticas como *The New York Times* (EUA)*, The Guardian* (Reino Unido), *El País* (Espanha), Estadão (Brasil), entre outros.

Em uma investigação com um conjunto de visualizações, Segel e Heer (2010) definiram alguns aspectos que podem subsidiar na composição de narrativa com dados, como por exemplo, apresentação de *slides* em movimento e histórias com detalhamento, isto é, com profundidade informativa, cujo foco maior, foi equilibrar os dados com os aspectos interativos. Neste contexto de narrativas e propriedades de dados em sua construção, a investigação de Nora Paul (2012) estabelece tipologias para as narrativas digitais, a saber: 1) Mídia: suporte da narrativa que combina recursos digitais como texto, fotos, gráficos, animação); 2) Ação: refere-se a dois aspectos distintos das narrativas digitais: o movimento do próprio conteúdo e a ação requerida pelo usuário para acessar o conteúdo; 3) Relacionamento: está ligado ao relacionamento entre o usuário do conteúdo e o próprio conteúdo. 4) Contexto: Capacidade de ofertar dados extras, remetendo a outros materiais, a exemplo dos *links*. e 5) Comunicação: Capacidade de conectar com outros tipos de mídias.

A pesquisa partiu da questão: Como as narrativas guiadas por dados endereçam formatos inovadores para a visualização de dados? Portanto, pretendemos observar e caracterizar esses elementos agregados às visualizações de dados e *Big data* visando compreender a perspectiva no contexto aqui explorado.

2 OBJETIVOS

Esta investigação tem como objetivo compreender o processo de utilização dos elementos narrativos que estão presentes nas visualizações de dados, realizando um estudo exploratório, a partir das categorias de análise: Elementos gráficos, interação, Dados estruturados, Proveniência dos dados, Tecnologias usadas e Inovação.

3 PROCEDIMENTOS METODOLÓGICOS

Para atender o objetivo deste estudo, procedemos com um Estudo de Caso e nos baseamos na literatura específica além do aspecto empírico da observação e análise dos recursos narrativos nas visualizações de dados vencedoras do *Data Journalism Awards 2018*[1], evento que acontece anualmente desde 2012, e esse ano foi em Lisboa, Portugal, e premia os melhores trabalhos com dados em jornais, revistas e agências de notícias mundiais[2]. Em 2018, foram 630 trabalhos submetidos ao prêmio de 58 países. Integralizando um total de 86 finalistas, com 13 trabalhos vencedores. Partimos para nossa pesquisa dessa amostra de 13 trabalhos e delimitamos um *corpus* de 8 visualizações que atendiam ao recorte ou escopo de jornalismo de dados. Descartamos do nosso estudo as produções que não se adequam à nossa amostra, como *startup*, *app*, *sites* e portfólios. O procedimento envolveu revisão de literatura, monitoramento e aplicação de Ficha de Observação[3] submetidas para essas visualizações no período de 17 a 26 de setembro de 2018. Consideramos as seguintes categorias de observação, quais sejam: **Elementos gráficos de visualização de dados** para Averiguar a presença de ilustrações, vídeos, gráficos estáticos, gráficos dinâmicos, áudios, textos curtos, fotos estáticas, fotos em movimentos, imagens em 3D, mapas dinâmicos ou estáticos, dentre outros, de modo que facilitem a narrativa; **Interação:** Aspecto fundamental nas visualizações, Cairo (2008) define três aspectos: Instrução (nível básico, quando a narrativa é linear e apresenta botões de avançar e retroceder); Manipulação (Poder mudar os objetos na visualização); e Exploração (Tipo de interação por imersão, ex. imagens de 360, com navegação hipertextual); **Dados Estruturados (ou semânticos)**: A partir dos estudos de Cairo (2011), essa categoria se refere em apresentar os dados de modo compreensível e com sentido a esse conjunto de dados, se está estruturado de modo compreensível a partir dos formatos (verticalizados, horizontalizados). "Também criamos informação estruturada quando desenhamos gráficos que codificam os números, o que de outra maneira seria impossível extrair algo útil" (CAIRO, 2011, p. 31, tradução nossa)[4]; **Proveniência dos dados:** Identificar a origem dos dados que foram explorados na visualização, se foram dados públicos, privados, de empresas, agências de notícias, etc. A ideia é identificar como esses dados estão sendo explorados; **Tecnologias usadas:** Identificar quais foram as tecnologias e linguagens mais utilizadas para a execução das produções de visualização de dados: E **inovação** como construto dessas narrativas em base de dados com estratégias de uso de novos recursos.

4 RESULTADOS

A partir dos dados coletados[5] da observação e do procedimento da análise dos mesmos, foi possível identificar que a utilização de elementos gráficos nas visualizações fazendo composição com os dados (exemplo de mapas, infografias, fotos e interações) para dar vazão a camada de dados e de construção de visualizações atrativas que possam dar sentido a dados complexos e cruzamento

1 Disponível em: https://goo.gl/TeKwkw Acesso em: 14 set 2018.
2 A competição é organizada pelo *Global Editors Network* e tem apoio do *Google News Initiative* e *Knight Foundation*.
3 Disponível em https://goo.gl/CBFr1w acesso em 3 out.2018
4 No original: "También creamos información estructurada cuando dibujamos gráficos que codifican los números, lo que de otra manera sería imposible extraer algo útil".
5 Disponível em https://goo.gl/HoVvAu acesso em 3 out. 2018.

de dados. Nos 8 casos analisados (Apêndice A), os mapas estáticos predominam com 87,5% como lugar para receber dados, com exceção do *PostMedia,* que não trouxe nem mapas estáticos e nem dinâmicos. Em seguida, o uso de vídeos enquanto recurso das narrativas com 75% de uso. Em compensação, tivemos 4 gráficos interativos em detrimento de 3 gráficos estáticos. Gráficos guiados por dados são 2 (*Infotimes* e *RunRun*), em outros momentos os dados vêm incorporados em outros elementos.

No aspecto interação, como recurso que permite maior participação do usuário no manejo desses dados e de toda a narrativa, identificamos produções mais avançadas e com uso mais intenso de características de interação com o predomínio do tipo de interação por Exploração, total de 100% dos casos analisados (8 de 8), além de 2 casos de Manipulação (*PostMedia* e *Infotimes,* que também concomitantemente apresentaram o grau de exploração, totalizando 25% para esse item). Nenhum caso foi registrado do nível mais básico, o de Instrução. Essa possibilidade de interação é um aspecto pertinente para o formato de visualização de dados porque quanto maior o grau de interação, mas aumenta a possibilidade de cruzamento de dados em nível de personalização pelo usuário. Logo, constitui-se de um recurso fundamental na construção atual dessas visualizações.

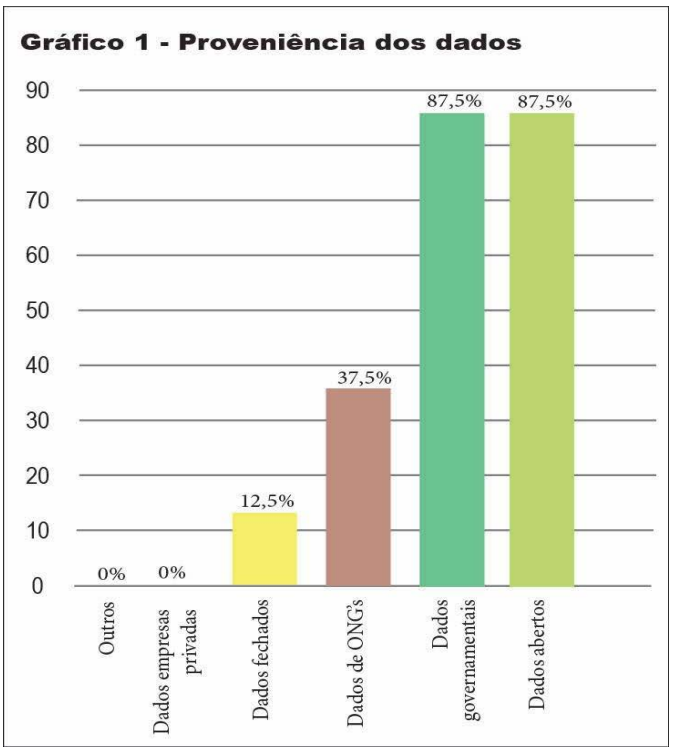
Em relação aos dados estruturados (Apêndice C), as narrativas de visualização de dados têm se sofisticado e utilizado novas possibilidades de apresentação. Identificamos a tendência de verticalização, totalizando 87,5% da amostra (7 casos) contra 2 casos (25%) horizontais. O modelo vertical se tornou tendência a partir do *SnowFall* do *The New York Times*, que a partir de 2012[6] iniciou o fenômeno da verticalização como formato para reportagens com dados permitindo aprofundamento para uso de recursos diversos como mapas, vídeos, mapas guiados por dados e outros recursos multimídia que dialogam diretamente com os dados em grande volume. A partir de então, os demais jornais mundiais começaram a aderir a esse formato, cuja "abundância do texto verbal sinaliza um resgate da qualidade, apuração e contextualização" (LONGHI; WINQUES, 2015, p. 8).

Além das narrativas verticais, os resultados apontaram para 75% (6 dos 8 casos analisados) das produções baseadas em mapas guiados por dados, cujos dados ficaram visíveis, seguidos de 25% do formato horizontalizados. Essas construções podem apontar tendência no jornalismo e na Ciência da Informação como formatos de apresentação. Esse tipo de formato também é denominado de narrativa *longform*, "um nível mais aprofundado de relato, que vai além do padrão cotidiano da produção e narrativas atraentes, frequentemente com elementos multimídia, que realçam o artigo" (LONGHI; WINQUES, 2015, p. 3).

As narrativas de visualização guiadas por dados se caracterizam, na origem das fontes, pelo uso de dados abertos (gráfico 1). Dos 8 casos analisados de visualização de dados, a maioria (7, ou seja, 87,5%) utilizam dados abertos. Igualmente o uso de dados governamentais (87,5%) e 3 casos de uso de dados de ONGs e similares (37,5%).[7] Dados fechados apenas no caso da agência *Reuters* (12,7%), que concomitantemente também utiliza dados abertos e governamentais nas suas produções. Portanto, nesse quesito os dados abertos são predominantes nos projetos de visualização de dados apresentados demonstrando avanço nesta política. Buscamos, nesse caso, compreender o avanço do uso de dados abertos e governamentais e sua abundância para construção de narrativas ou projetos.

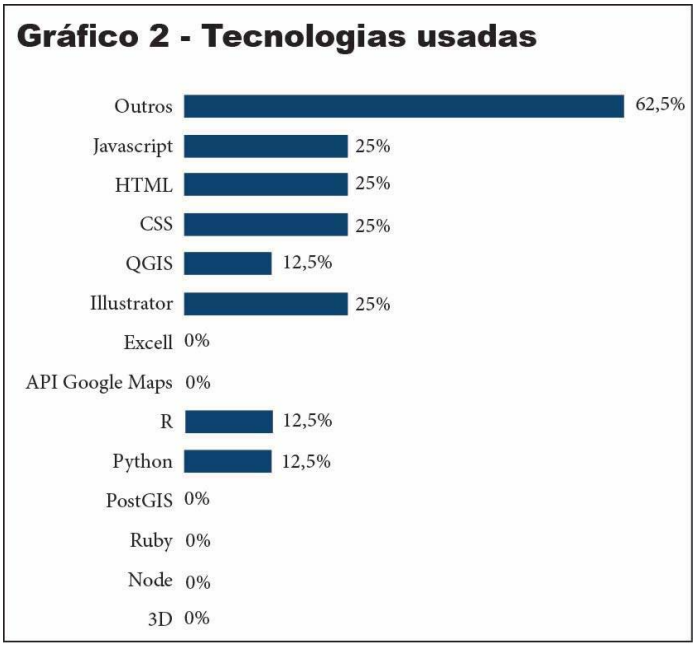
6 Disponível em:https://goo.gl/zUufx2 Acesso em: 18 set 2018.

7 Neste levantamento, os resultados somam mais de 100% porque era possível selecionar mais de um elemento por caso



Fonte: Dados da Pesquisa, 2018.

Sobre as tecnologias utilizadas (gráfico 2), esse item é mais complexo de analisar tendo em vista que as tecnologias e linguagens dependem do tipo de projeto e da complexidade do mesmo e, ao mesmo tempo, do tamanho e da especialidade da equipe envolvida em tais projetos. O predomínio na nossa observação foi de tecnologias fora do escopo do nosso levantamento prévio e enquadrando-se em "outros". Dos 8 casos, 5 utilizaram tecnologias distintas e novas (65,5%). Há uma variedade de tecnologias e linguagens observada na análise, a exemplo da linguagem *R* e *Python* com 12,5% do total. Outros como JavaScript, CSS, por exemplo, totalizaram 25%.



Fonte: Dados da Pesquisa, 2018.

Especificamos apenas as linguagens e tecnologias mencionadas no resumo executivo (apresentação) dos projetos vencedores sem exploração interna dos projetos para tentar identificar pelo código fonte ou outro meio tendo em vista que isso exigiria um trabalho de "bastidores" dos projetos de difícil caracterização visando a precisão. Logo, partimos para o próprio resumo em que cada projeto indicava essas tecnologias utilizadas, de modo a termos um recorte mais coerente. Essa exploração nos fornece um panorama e aponta para o uso crescente de tecnologias específicas de visualização de dados como o *Tableau Public*[8]. Na nossa observação, exploramos o conceito de inovação, partimos da observação de cada caso e dos próprios apontamentos no resumo executivo dos projetos para poder apontar os avanços em termos de inovação nessas narrativas baseadas em visualização guiada por dados. Nesta direção, identificou-se que todas, em algum aspecto, inclusive pela natureza do prêmio, apresentam alguns elementos de inovação e de avanço quanto ao estado das produções do momento, apresentando inovações disruptivas para os formatos de visualização de dados. De um modo geral, podemos apontar alguns aspectos inovadores como: 1) Dados Geolocalizados por satélite e atualizáveis em tempo real; 2) Mapas guiados por dados como bússola para as informações; 3) Imagens em 360 grau conferindo imersão na narrativa e 4) *Machine Learning* e Inteligência Artificial como potencializadores dessas narrativas.

5 CONSIDERAÇÕES FINAIS

As narrativas em ambientes digitais criaram novos paradigmas em relação às narrativas tradicionais, tendo em vista o potencial da ambiência digital para a construção das mesmas. Nesta direção, esse estudo tentou compreender os processos de construção das narrativas com dados nestas visualizações e demarcar as estratégias utilizadas como forma de compreensão dos dados. Observamos alguns aspectos que servem para reflexão: 1) Formato *Longform* segue como uma prática que vem sendo cada vez mais explorado nas narrativas com dados; 2) Mapas guiados por dados como estrutura principal de apresentação dos dados, o que confere um aspecto de inovação dos dados; 3) Narrativas com dados em profundidade, o que quebra com o paradigma de somente mostrar, mas contextualizar, dar sentido àqueles dados visualmente.

Os textos curtos e blocados ao lado das imagens foram bem recorrentes, assim como o uso dos *hiperlinks*. 4) Dados abertos cada vez mais sendo explorados nestas produções, o que pode transparecer uma certa ética ao creditar nas narrativas, muito embora não se sabe se todos os dados brutos foram trabalhos massivamente. Vale destacar a imagem em satélite como elemento diferencial produzindo um ambiente de imersão naquela estrutura de dados, ainda que só a *Reuters* tenha utilizado o recurso timidamente. As narrativas seguem as estratégias de tendências de mercado que estão postas, como formatos *Longform* e imagens de satélite de modo imersivo, para que se sinta dentro da narrativa, muito embora apenas a *Agência Reuters* utilizou, mas segue sendo um indício promissor desse recurso. A opção pela interação por exploração, cuja navegação se deu mais pela força do *link*, condiz com a proposta da narrativa, de contextualizar os dados, contar a história dos acontecimentos, o que pode parecer pouco dinâmico à primeira vista. Em nossa análise, alguns

8 Software livre que permite criar visualizações de dados na Web. Disponível em: https://public.tableau.com/pt-br/s/ Acesso em: 1 out 2018.

veículos analisados poderiam explorar mais imagens em 360 graus, por exemplo, como ferramenta imersiva das narrativas, o que inovaria no manejo com os dados.

É válido ressaltar que não há modelos fechados para realizar análises e estudos, mas as existências de vários ângulos de abordagens em diversas áreas, tendo em vista que "a narrativa é capaz de corresponder a uma cronologia cronológica e não cronológica, extraindo uma sucessão de eventos" (BOASE, 2013, p. 5)[9]. Assim sendo, a pesquisa avança no sentido de demarcação de como esses elementos narrativos estão sendo explorados e seus potenciais na era do *Big Data*, de modo em o processo analítico-visual seja um dos expoentes, e ainda que resguardando as limitações do estudo. Para trabalhos futuros, busca-se realizar um estudo de recepção a fim de averiguar como os usuários recebem e compreendem esses dados a partir dessas visualizações, bem como a exploração mais detalhada das tecnologias utilizadas pelos profissionais.

REFERÊNCIAS

BOASE, C. Digital **Storytelling for Reflection and Engagement:** a study of the uses and potential of digital storytelling. Centre for Active Learning & Department of Education, University of Gloucestershire. (2013) Disponível em: <https://goo.gl/MNJg9H>. Acesso em 3 set 2018.

CAIRO, A. **El arte funcional: infografía y visualización de la información.** Madrid: Alamut, 2011._____. (2008) Infografía 2.0: visualización interactiva de información en prensa. Alamut, Madrid.

FIGUEIRAS, A. A typology for data visualization on the web. In Information Visualisation (IV), 2013 **17th International Conference,** pages 351–358, July 2013. Disponível em: <https://goo.gl/TNhtEm>. Acesso em: 18 mai 2018.

KOSARA, R; MACKLINLAY, J. Storytelling: The next step for visualization. **Compute**r, 46(5):44–50, 2013.

LAGOZE, C. Big Data, data integraty, and the fracturing of the zone control. **Big Data & Society.** jul./dec. 2014. Disponível em: <https://goo.gl/TcsW1r>. Acesso em: 10 abr. 2018.

LONGHI, R.; WINQUES, K. O lugar do longform no jornalismo online: qualidade versus quantidade e algumas considerações sobre o consumo. Brasília. 24º Compós2015, - **Anais..** Brasília: Universidade de Brasília, 2015. v. 1. p. 1-19.

PAUL, N. Elementos das narrativas digitais. In P. Ferrari (Org). **Hipertexto e hipermídia:** as novas ferramentas da Comunicação digital. São Paulo: Contexto, 2012.

SEGEL, E; HEER, J. Narrative visualization: Telling stories with data. **Visualization and Computer Graphics, I**EEE Transactions on, 16(6):1139–1148, 2010. Disponível em: <https://goo.gl/TmNgFu>. Acesso em: 3 set 2018.

9 No original: "Narrative is able to combine both a chronological and non-chronological dimension into a meaningful whole by extracting a configuration from a followable succession of events".

TUFTE, E. **The visual display of quantitative information**. Cheshire, Connecticut: Graphic Press, 2001.

VIÉGAS, F. **Designer explica como a visualização de dados pode ser atraente**. Rio de Janeiro, TV Globo, 20 jun. 2013. Entrevista a Globo Universidade. Disponível em: <http://goo.gl/c35Ztn>. Acesso: 4 jul 2017.

Apêndice A – Elementos gráficos nas visualizações

TIPO	Agência Reuters	La Nación	PostMedia	Infotimes	Caixin Media	RunRun	BuzzFeed	G1	TOTAL
Gráficos interativos	X		X	X	X				4 (50%)
Gráficos guiados por dados				X		X			2 (25%)
Gráficos estáticos	X				X	X			3 (37,5)
Mapas interativos	X			X	X	X			4 (50%)
Mapas estáticos	X	X		X	X	X	X	X	7 (87,5%)
Foto interativa	X								1 (12,5%)
Foto estática	X			X		X	X		4 (50%)
Imagem em 360	X								1 (12,5%)
Ilustrações	X		X	X	X		X		5 (62,5%)
Vídeo		X		X	X	X	X	X	6 (75%)
Imagem de satélite	X								1 (12,5%)
Infográficos estáticos								X	1 (12,5%)
Infográficos interativos									0 (0%)
Outros			X						1 (12,5%)

Fonte: Dados da Pesquisa, 2018.

Apêndice B – Interação

TIPO	Agência Reuters	La Nación	PostMedia	Infotimes	Caixin Media	RunRun	BuzzFeed	G1	TOTAL
Instrução (linear e apareenta botões de avançar e retroceder)									0 (0%)
Manipulação (quando é possível mudar os objetos dentro da visualização)			X	X					2 (25%)
Exploração (tipo de interação com navegação hipertextual, explorando dados e/ou por imersão)	X	X	X	X	X	X	X	X	8 (100%)

Fonte: Dados da Pesquisa, 2018.

Apêndice C – Dados estruturados

TIPO	Agência Reuters	La Nación	PostMedia	Infotimes	Caixin Media	RunRun	BuzzFeed	G1	TOTAL
Verticalizados	X	X	X	X	X		X	X	7 (87,5%)
Horizontalizados						X	X		2 (25%)
Mapas guiados por dados	X	X		X	X	X		X	6 (75%)
Outro									0 (0%)

Fonte: Dados da Pesquisa, 2018.

O ENSINO DA E-SCIENCE NO ÂMBITO DOS PROGRAMAS BRASILEIROS DE PÓS- GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

THE TEACHING OF E-SCIENCE IN THE CONTEXT OF BRAZILIAN GRADUATE PROGRAMS IN INFORMATION SCIENCE

Renata Lemos dos Anjos[1], Débora Gomes de Araújo[2]
(1) Universidade Federal da Paraíba, Programa de Pós-Graduação em Ciência da Informação, renatalemosdosanjos@gmail.com.
(2) Universidade Federal da Paraíba, Programa de Pós-Graduação em Ciência da Informação, debora.g.de.araujo@gmail.com.

Resumo:
O constante avanço das Tecnologias Digitais de Informação e Comunicação provocou mudanças no cenário científico, alterando o modo de pensar e fazer ciência. O quarto paradigma da ciência, denominado como e-*Science*, pode ser definido de forma resumida como uma ciência amplamente orientada a dados oriundos de investigações científicas. Esta realidade mostra ser um amplo campo de atuação para os profissionais da Ciência da Informação. A partir de uma pesquisa exploratória e documental, com análise quantitativa e qualitativa, investigamos se as disciplinas componentes das grades curriculares dos Programas de Pós-Graduação em Ciência da Informação avaliados e reconhecidos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, abordam sobre a ampla realidade da e-*Science,* suas variações e práticas. Diante dos resultados obtidos com a pesquisa, obtivemos que dentre os vinte e três Programas de Pós-Graduação na área da Ciência da Informação, onze, representando 47,82% dos programas oferecem disciplinas que versam sobre o fenômeno da e-*Science*, suas relações, suas variações e suas práticas. Percebeu-se que os Programas de Pós- Graduação em Ciência da Informação analisados estão se adequando, bem como, apresentando cursos pioneiros nesta realidade da e-*Science*, como o curso de mestrado profissional em Gestão de Dados de Pesquisa oferecido pela Universidade Federal do Estado do Rio de Janeiro.
Palavras-chave: e-Science. Ciência da Informação. Pós-Graduação em Ciência da Informação.

Abstract:
The constant advance of the Digital Information and Communication Technologies caused changes in the scientific scene, transforming the way of thinking and doing science. The fourth paradigm of science, called e-Science, can be briefly defined as a science largely geared to data from scientific research. This reality shows to be a broad field of action for the Information Science Professionals. From an exploratory and documentary research, with quantitative and qualitative analysis, we investigated whether the disciplines that are part of the curricula of the Graduate Programs in Information Science evaluated and recognized by the Coordination of Improvement of Higher Education Personnel, are within the reality of e-Science, its variations and practices. Considering the results obtained with the research, we found that among the twenty-three Graduate Programs in the area of Information Science, eleven, representing 47.82% of the programs offer subjects that deal with the phenomenon of e-Science, its relations, its variations and its practices. It was noticed that the Graduate Programs in Information Science analyzed are adequate, as well as presenting pioneering courses in the context of e-Science, such as the professional master's degree in Data Management Research offered by the Federal University of the Rio de Janeiro state.
Keywords: e-Science. Information Science. Graduates studies in Information Science.

1 INTRODUÇÃO

O constante avanço das Tecnologias Digitais de Informação e Comunicação – TDIC provocou mudanças no modo de pensar e fazer ciência, causando rupturas e alterações de paradigmas seguidos em todo o domínio do fazer científico. Shuttleworth (2008, online, tradução nossa) afirma que "uma mudança de paradigma não é uma ameaça à ciência, mas sim a própria maneira pela qual ela progride[1]". Desta forma, paralelamente ao desenvolver da tecnologia, o modo de pensar e fazer ciência foi sofrendo alterações, até que estes momentos de rupturas foram tidos como paradigmas da ciência, configurando-se como uma evolução da mesma.

De acordo com Gray (2007), os paradigmas da ciência correspondem a fases, momentos do empreendimento científico e notoriamente às fases da sociedade de uma maneira geral. Inicialmente, há mil anos, o primeiro paradigma da ciência, nomeado como "Ciência Experimental" (HEY et al., 2009), a ciência acontecia de forma exclusivamente experimental, empírica, por meio da observação e descrição das leis naturais.

No que concerne ao segundo paradigma da ciência, o da "Ciência Teórica", por volta do século XVII, ela passa a basear-se em modelos teóricos, como as Leis de Kepler, Leis de Newton, Equações de Maxwell, assumindo um ramo teórico (GRAY, 2007; HEY et al., 2009).

Mais recentemente, na segunda metade do século XX, esses modelos teóricos tornaram-se complicados em sua resolução, dando início a simulações e ao uso de recursos computacionais, reconhecido como o terceiro paradigma da ciência, a "Ciência Computacional" (HEY et al., 2009). Esta abordagem faz uso da computação tanto para a simulação de fenômenos complexos, quanto para solucionar equações demandavam um árduo trabalho pesquisadores anterior ao uso computadores (GRAY, 2007).

Em seu terceiro paradigma, a ciência passou a produzir uma quantidade massiva de dados, por meio do uso intensivo da infraestrutura computacional disponível, progredindo para o mais atual paradigma da ciência, o quarto paradigma.

No quarto paradigma da ciência, esses dados passam a ser explorados e considerados "produtos valiosos do empreendimento científico[2]" (STRASSER et al, 2012, p. 2, tradução nossa). Desta forma passam a ser considerados como dados científicos primários que acabam por serem utilizados como matéria-prima para futuras investigações científicas. Com base nesta realidade, Tenopir (2011, p. 1, tradução nossa) afirma que "seguindo os paradigmas de investigações anteriores, esta nova era tem sido chamada de 'o quarto paradigma: descoberta científica do uso intensivo dos dados[3]'", também definido como e-Science.

A e-Science pode ser definida de forma resumida por Gray (2005, apud Sales et al., 2013) como uma "ciência produzida a partir do uso, processamento, análise e compartilhamento de dados de pesquisa". Ou seja, uma ciência amplamente cumprida e realizada nas bases do quarto paradigma científico, o qual valoriza abundantemente os dados oriundos de investigações científicas.

Nesta perspectiva, percebe-se que a e- Science configura-se por ser uma ciência baseada sob a construção de uma infraestrutura informática computacional de uso distribuído, a qual permite

1 A paradigm shift is not a threat to science, but rather the very manner in which it progresses.
2 Research data are valuable products of the scientific enterprise (...).
3 Following the previous research paradigms, this new era has been called "the fourth paradigm: data-intensive scientific discovery.

e incentiva a colaboração à distância entre os pares, entre equipes de pesquisa, envolvendo a todo instante o uso intensivo e o constante compartilhamento de dados e recursos computacionais (ALBAGLI et al., 2013).

Desta forma, a Ciência da Informação depara-se com o fenômeno da e-Science, para com o qual, pode e deve contribuir para que se desenvolva da melhor maneira. Ao identificar e estudar fatores e características que vão propiciar o equilíbrio entre os atores envolvidos no processo e a máxima otimização do uso e reuso dos dados (SANT'ANA, 2016), além de revelar-se como um amplo campo de atuação para os profissionais da informação.

Diante da perspectiva de que, está diretamente associado aos processos de coleta, curadoria, processamento, armazenamento, análise e visualização dos dados, e que as tendências das pesquisas na CI, neste contexto, versam sobre a "proveniência, as melhores práticas, o planejamento, a curadoria entre outras atividades cientificas relacionadas à teoria e à práxis com dados de pesquisa" (SEMELER, 2017, p. 137). Percebe-se deste modo, que a CI está amplamente inserida no âmbito do quarto paradigma científico, tanto em questões de investigações científicas, quando na atuação dos profissionais da informação.

Neste contexto, com base no fenômeno emergente da e-Science, indagamos se a realidade da e-Science é abordada nos componentes curriculares dos Programas de Pós-Graduação em Ciência da Informação avaliados e reconhecidos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)?

2 OBJETIVO

A pesquisa em tela busca desvendar o ensino da e-Science no âmbito dos programas brasileiros de Pós-Graduação em Ciência da Informação avaliados e reconhecidos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

3 PROCEDIMENTOS METODOLÓGICOS

O presente estudo caracteriza-se como uma pesquisa exploratória e documental, com análise quantitativa e qualitativa.

Inicialmente, foram identificados vinte e três Programas de Pós-Graduação em Ciência da Informação avaliados e reconhecidos pela CAPES na Plataforma Sucupira, bem como, os programas que ofereciam cursos a nível de mestrado acadêmico, mestrado profissional, e doutorado. Ressalta-se que foram contabilizados os programas indexados pela CAPES na área de avaliação "Comunicação e Informação" e posteriormente aqueles da área do conhecimento "Ciência da Informação", totalizando em vinte e três programas.

Após a identificação dos Programas de Pós-Graduação em Ciência da Informação, partimos para uma visita aos sites/portais buscando as disciplinas, obrigatórias e optativas, componentes curriculares de cada programa. Ressaltamos que ao não encontrarmos as grades curriculares dos programas, buscamos pela oferta de disciplinas relativas aos semestres 2018.1 e 2018.2, e também

na própria plataforma Sucupira, na qual os programas devem informar à plataforma as disciplinas que são oferecidas.

Ao dar seguimento a pesquisa, percebeu-se que algumas das disciplinas identificadas abordam temas ligados a e- Science, tais como, big data, humanidades digitais (digital humanities), tratamento de dados em geral, bem como, Data Science, uma outra nomenclatura para um fenômeno que pode estar classificado no contexto investigado, as quais foram contabilizadas para esta pesquisa.

Ressalta-se que, também foram avaliadas as ementas das disciplinas, quando disponíveis, em virtude de identificação de relações com o tema proposto.

Contabilizamos que dentre os vinte e três Programas de Pós-Graduação em Ciência da Informação identificados, onze dispõem em suas componentes curriculares de disciplinas que versam e preparam seu corpo discente para a realidade do quarto paradigma científico.

4 RESULTADOS

Após realizada a busca pelos componentes curriculares em cada site/portal dos programas selecionados, e também na Plataforma Sucupira, obtivemos que onze programas oferecem disciplinas com conteúdos aderentes ao contexto da e-Science. O Quadro 1 apresenta estas disciplinas.

Quadro 1: A e-Science nos Programas de Pós-Graduação em Ciência da Informação.

PROGRAMA	DISCIPLINAS
PPGCI - UNESP	- A Ciência da Informação e o ciclo de vida dos dados: fronteiras de pesquisa; - Aspectos tecnológicos de acesso à dados abertos para Ciência da Informação; - Conceitos e tecnologias para publicação de dados abertos e semânticos seguindo as melhores práticas do Linked Data; - Design da informação e curadoria digital; - Fundamentos tecnológicos de representação de dados e informações; - Web semântica: conceitos e tecnologias.
PPGCIN - UFSC	- Big Data e Web Semântica. - Humanidades digitais e Ciência da Informação; - Tendências tecnológicas aplicadas a dados.
PPGCI - UFPB	- Big Data e a Sociedade da Informação; - e-Science no âmbito da Ciência da Informação; - Gestão de Dados Científicos na Perspectiva da e-Science.
PPGGOC - UFMG	- Metadado: formatos e modelos; - Mineração e modelagem de dados.
PPGCINF - UNB	- Ciência de dados aplicada a curadoria de dados científicos; - Ciência de dados aplicada – modelos de análise.
PPGCI – ECA/USP	- Humanidades digitais.
PPGCI - UEL	- A mineração de dados aplicada à pesquisa social.
PPGCI - UFPE	- Curadoria digital.
PPGCI - UFS-Car	- Representação e metadados em ambientes digitais.
PPGCI - UFF	- Acesso aberto à informação científica;
PPGB - UNI-RIO	- Gestão de dados de pesquisa (MP).

Fonte: Dados da Pesquisa, 2018.

Com base no exposto, verificou-se que aproximadamente 47,82% (onze) dos programas dispõem de disciplinas que versam sobre a realidade da e-Science, suas relações e variações.

O Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual de São Paulo (PPGCI/UNESP), foi identificado como o programa mais atualizado em questões que permeiam a e-Science. Oferecendo tanto disciplinas teóricas quanto disciplinas práticas sobre essa realidade.

Após o PPGCI/UNESP, obtivemos uma situação de equivalência entre os Programas de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina (PPGCI/UFSC), e o da Universidade Federal da Paraíba (PPGCI/UFPB) os quais oferecem três disciplinas nesta seara, ambientando seus discentes no contexto da e-Science. O PPGCI/UFSC, aborda a e-Science de uma forma geral por meio de três disciplinas que versam sobre temas que relacionam-se com a mesma, conforme apresentado no Quaro 1.

No que tange ao PPGCI/UFPB, identificamos que o mesmo oferece disciplinas que versam sobre o Big Data e a Sociedade da Informação – relacionados com a e-Science; além de também oferecer uma disciplina que aborda o fenômeno da e-Science no âmbito da CI; e uma outra que trata sobre a gestão dos dados científicos, revelando a afinidade entre o quarto paradigma científico e a CI.

Obtivemos uma outra situação de equivalência entre o Programas de Pós-Graduação em Gestão & Organização do Conhecimento da Universidade Federal de Minas Gerais (PPGGOC/UFMG) e o Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília (PPGCI/UNB), os quais oferecem duas disciplinas nesta vertente conforme mostrado no Quadro 1. O PPGGOC/UFMG, oferece disciplinas de cunho aplicado, como o uso dos metadados para descrição acurada dos conjuntos de dados provenientes das pesquisas, como também, práticas de mineração e modelagem de dados utilizadas no tratamento dos dados científicos. E o PPGCI/UNB oferece duas disciplinas na vertente da e-Science, sendo por muitos definida como a Ciência de dados, por ser toda orientada a dados.

A Escola de Comunicações e Artes da Universidade de São Paulo, oferece a Pós-Graduação em Ciência da Informação (PPGCI/ECA-USP), a qual oferece a disciplina "Humanidades Digitais" que se refere a uma área de pesquisa e ensino que versa sobre a intersecção das tecnologias para com às ciências humanas, baseando-se nas práticas de mineração de dados, recuperação da informação, e visualização digital dos produtos gerados pelas ciências, o que a e-Science também orienta (SEGUNDO, 2015).

Já o Programa de Pós-Graduação em Ciência da Informação da Universidade de Londrina (PPGCI/UEL), oferece a disciplina "Mineração de dados aplicada à pesquisa social" apresentando-se como uma disciplina que capacita os profissionais da informação no trabalho da mineração de dados, tão utilizado neste momento do quarto paradigma científico.

O Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco (PPGCI/UFPE), oferece a disciplina de Curadoria Digital a qual versa sobre determinados subtemas: ciclos de curadoria digital; preservação da informação em ambientes digitais; metadados de preservação digital; modelos e estratégias de preservação digital; gestão de riscos em preservação digital; e-Science. Desta forma, além de abordar de forma geral a prática da curadoria digital, que "são procedimentos práticos e teóricos" (SAYÃO; SALES, 2016, p. 68), também versa sobre a realidade da e-Science.

A Universidade Federal de São Carlos em seu Programa de Pós-Graduação em Ciência da Informação (PPGCI/UFSCar), oferece uma disciplina prática de representação e metadados em ambientes digitais, a qual pode ser utilizada no contexto da e-Science.

No Rio de Janeiro, a Universidade Federal Fluminense, oferece o Programa de Pós-Graduação em Ciência da Informação (PPGCI/UFF), que versa sobre uma das vertentes da e-Science, que é o acesso aberto a informações/dados científicos, uma prática amplamente incentivada em tempos do quarto paradigma científico.

Ainda no Rio de Janeiro, e apresentando novidades na área, a Universidade Federal do Estado do Rio de Janeiro (UNIRIO), lançou recentemente, o edital de seleção para um mestrado profissional, o qual terá como temática "Gestão de Dados de Pesquisa[4]", tornando-se uma iniciativa pioneira no oferecimento de curso em pós-graduação que aborda amplamente as práticas incentivadas pela e-Science, por meio das três disciplinas, que foram divulgadas até o momento.

A disciplina aspectos contemporâneos da comunicação científica, versará sobre o movimento do acesso aberto, ciência aberta, dados abertos, inovações na comunicação científica, entre outros temas, os quais estão diretamente relacionados à realidade da e-Science.

Uma outra disciplina é a de inovação aberta e ciência aberta, que discorrerá sobre as definições e conceitos de invenção, design, pesquisa, desenvolvimento tecnológico e inovação; sobre as dimensões da inovação tecnológica, humana, econômica, organizacional, social e outras; inovação aberta: conceito, métodos e práticas; apresentará a relação entre inovação aberta e ciência aberta; gestão, plataformas abertas e ecossistema de inovação.

Por fim, a disciplina gestão de dados de pesquisa, a qual abordará os plano de gestão de dados; os benefícios e desafios no compartilhamento dos dados; as práticas de citação dos dados, de proteção e confidencialidade dos dados; sobre propriedade intelectual, autenticidade e integridade dos dados; curadoria de dados e boas práticas como armazenamento, backup, segurança, arquivamento e preservação dos dados; também abordará os princípios FAIR e como executá-los por meio da estrutura GOFAIR, além do papel e responsabilidade dos bibliotecários na gestão de dados de pesquisa. Estando expressamente relacionada à realidade e emergência do quarto paradigma científico.

4 CONCLUSÃO OU CONSIDERAÇÕES FINAIS

Ressalta-se que o PPGCI/UNESP foi identificado como o programa que mais adequa-se a esta realidade. Também se verificou indícios de que tanto a CI quanto os profissionais da informação relacionam-se diretamente com a e-Science, ciência dos dados. Nesta perspectiva, a UNIRIO abriu seleção para um novo curso de pós-graduação, o mestrado profissional em Gestão de Dados de Pesquisa. O que nos possibilita observar que os programas estão se atualizando no que concerne a oferta de disciplinas, e começam a ter iniciativas pioneiras nesta realidade do quarto paradigma científico.

4 http://www.unirio.br/ppgb/arquivo/editais-de-selecao/edital-tematico-2018-gestao-de-dados-de-pesquisa

REFERÊNCIAS

ALBAGLI, S. *et al*. Ciência Aberta: correntes interpretativas e tipos de ação. **Liinc em Revista**, v. 10, n. 2, p. 434 – 450, 2014. Disponível em: <http://revista.ibict.br/liinc/article/view/3593>. Acesso em: 24 set. 2018.

GRAY, J. **eScience**: a transformed scientific method. Palestra apresentada no Conselho Nacional de Pesquisa dos Estados Unidos (NRC-CSTB). Mountain View, Califórnia, 2007.

HEY, T. *et al*. **The fourth paradigm: data- intensive scientific discovery**. Redmond, WA: Microsoft Research, 2009.

SALES, L. F. *et al*. Publicações ampliadas: um novo modelo de publicação acadêmica para o ambiente de e-Science. In: XIV Encontro Nacional de Pesquisa em Ciência da Informação, 2013, Florianópolis. **Anais...** Florianópolis: XIV ENANCIB, 2013. Disponível em: <http://eprints.rclis.org/20666/1/APRESENTA ÇÃO%20-%20ENANCIB%202013.pdf >. Acesso em: 24 set. 2018.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, v. 21, n. 2, p. 116–142, dez. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/infor macao/article/view/27940/20124>. Acesso em: 31 jul. 2018.

SAYÃO, L. F.; SALES, L. F. Curadoria digital e dados de pesquisa. **AtoZ**: novas práticas em informação e conhecimento, v. 5, n. 2, p. 67-71, dez. 2016. ISSN 2237-826X. Disponível em: <https://revistas.ufpr.br/atoz/article/view/4970 8/30179https://revistas.ufpr.br/atoz/article/view/49708/30179>. Acesso em: 25 set. 2018.

SEGUNDO, J. E. S. Web semântica, dados ligados e dados abertos: uma visão dos desafios do brasil frente às iniciativas internacionais. **Encontro Nacional de Pesquisa em Ciência da Informação**, v. 16, 2015. Disponível em: <http://www.brapci.inf.br/index.php/article/do wnload/43838>. Acesso em: 25 set. 2018.

SEMELER, A. R. **Ciência da informação em contextos de e-science**: bibliotecários de dados em tempos de Data Science. 2017. 164f. Tese (Doutorado) - Programa de Pós-

Graduação em Ciência da Informação, Universidade Federal de Santa Catarina, Florianópolis, 2017. Disponível em: <https://repositorio.ufsc.br/handle/123456789/ 185593>. Acesso em: 26 set. 2018.

SHUTTLEWORTH, M. What Is a Paradigm?. **Philosophy of Science, Explorabe.com**, Oct. 24, 2008. Disponível em: <https://explorable.com/what-is-a-paradigm>. Acesso em: 26 set. 2018.

STRASSER, C. *et al*. **Primer on data management**: what you always wanted to know. 2012. Disponível em: <https://www.dataone.org/sites/all/documents/D ataONE_BP_Primer_020212.pdf>. Acesso em: 27 set. 2018.

TENOPIR, C. *et al*. Data Sharing by Scientists: Practices and Perceptions. **Plos One**, v. 6, n.6, p. 1-21, 2011. Disponível em: <http://journals.plos.org/plosone/article?id=10. 1371/journal.pone.0021101>. Acesso em: 27 jul. 2018.

O PAPEL DO BIBLIOTECÁRIO NO PROCESSO DE RECUPERAÇÃO DA INFORMAÇÃO

THE ROLE OF THE LIBRARIAN IN THE INFORMATION RECOVERY PROCESS

Miriam Cândida de Jesus [1], Nina Cláudia Mendonça Campos de Miranda [2], Célia da Consolação Dias [3]
(1) UFMG, Av. Antônio Carlos, 6627, Campus Pampulha, Belo Horizonte – MG, miriamcandidadejesus@gmail.com
(2) UFMG, Av. Antônio Carlos, 6627, Campus Pampulha, Belo Horizonte – MG, ninacm.nm@gmail.com
(3) UFMG, Av. Antônio Carlos, 6627, Campus Pampulha, Belo Horizonte – MG, celiadias@gmail.com

Resumo:
Este trabalho aborda o papel do profissional bibliotecário no processo de recuperação da informação no ambiente de bibliotecas universitárias. O objetivo principal deste estudo é revisar na literatura o papel que o bibliotecário exerce como mediador informacional na educação e no desenvolvimento da competência informacional do usuário no processo de recuperação da informação. A proposta é fazer um link entre as competências desejáveis do profissional da informação e sua influência no processo de recuperação da informação pelo usuário. Para tanto, foi feita uma busca na literatura sobre a atividade de auxílio/orientação aos usuários, visando torná-lo usuário competente na recuperação da informação. O conhecimento do que a literatura recomenda no tocante às habilidades profissionais e o que se espera do bibliotecário pode contribuir para a adoção de uma postura mais ousada do bibliotecário, indo de encontro às expectativas do usuário no que se refere à recuperação da informação.
Palavras-chave: Recuperação da informação. Papel do bibliotecário. Usuário.

Abstract:
This work addresses the role of the professional librarian in the process of information retrieval in the university library environment. The main objective of this study is to review in the literature the role that the librarian exercises as an informational mediator in education and in the development of user informational competence in the information retrieval process. The proposal is to link the desirable competencies of the information professional and their influence on the process of information retrieval by the user. To do so, a search was made in the literature on the activity of assistance / guidance to users, aiming to make it a competent user in the information retrieval. Knowledge of what literature recommends regarding professional skills and what is expected of the librarian can contribute to the adoption of a bolder librarian's posture, meeting the expectations of the user with regard to information retrieval.
Keywords: Information retrieval. Librarian role. User.

1 INTRODUÇÃO

As bibliotecas universitárias devem estar atentas quanto à explosão informacional e ao avanço tecnológico a fim de contribuírem para o pleno desenvolvimento da habilidade do usuário em recuperar a informação obtendo os melhores resultados, em tempos de recursos eletrônicos e internet.

Esta nova realidade sugere que o usuário desenvolva, aperfeiçoe e use cada vez mais suas habilidades em buscar e usar a informação. Tais habilidades os capacitam não somente a aprender de forma autônoma e a construir seu próprio conhecimento, mas, sobretudo a dar suporte ao usuário na busca por informação.

Neste contexto a oferta de serviços de informação e a atuação educacional e de suporte do bibliotecário na biblioteca universitária merecem destaque. Este destaque está relacionado ao desafio de desenvolver mecanismos, serviços e atividades visando promover à independência e autonomia de seus usuários, tanto na identificação das fontes de informação quanto ao acesso e uso destes recursos.

Sadi *et al* (1985) enfatiza que, nas últimas décadas, as bibliotecas deixaram de ser apenas depositárias do conhecimento e passaram a ter um aspecto mais ativo, ou seja, suas atividades são desenvolvidas com foco no usuário, desenvolvendo sistemas de informação e outros serviços visando sua satisfação.

Observou-se muitas vezes que os usuários desconhecem a existência de fontes de informação disponíveis em meio eletrônico ou impresso, bem como as formas e técnicas de pesquisa.

Segundo Belluzo e Macedo (1993), as bibliotecas universitárias têm a necessidade de acompanhar a evolução das informações e das tecnologias, oferecendo melhor qualidade nos seus serviços. A biblioteca, como uma unidade de informação, se preocupa com o aperfeiçoamento de suas atividades, investe no desenvolvimento tanto dos seus recursos de informação quanto na sua equipe interna, bem como na implementação e no uso de tecnologias de informação que possibilita o acesso, a busca e a recuperação de informação pelos usuários.

Neste sentido, a biblioteca universitária desempenha um papel relevante como gestora e mediadora atuando na seleção, organização, disponibilização e recuperação de informação para dar suporte à geração de novos conhecimentos produzidos para o tripé que sustenta a universidade, isto é, o ensino, a pesquisa e a extensão. Para atender as inúmeras demandas a partir deste tripé cabe à biblioteca universitária lidar com recursos de informação que podem ser caracterizados como híbridos, formados por objetos informacionais de uma diversidade muito grande e que varia desde o acervo tradicional, aos objetos digitais, imagéticos, sonoros até os dados digitais, entre outros. A gestão destes recursos de informação para atender aos objetivos tão amplos das universidades demanda um grande esforço das bibliotecas universitárias e dos seus bibliotecários.

Com o foco no usuário, o papel do bibliotecário é de extrema relevância. Esta relevância está respaldada na ampliação da atuação do bibliotecário, sujeito mediador, que conhece as fontes de informação e responsável por criar estratégias para que o usuário possa conquistar autonomia na direção da construção do próprio conhecimento.

O termo mediação deve ser usado em substituição ao termo intermediário, tal como apontado por Kuhlthau (1994). Para a autora a mediação pressupõe uma interação humana entre

os envolvidos no processo de busca da informação e o intermediário intervém entre a informação e o usuário sem haver interação entre eles. A interação é justamente o que pode ser proposto pela biblioteca universitária no que diz respeito ao à oferta de serviços de informação e a disponibilização dos recursos.

Além deste papel de mediador apontado anteriormente por Kuhlthau (1994), outra função atribuída ao bibliotecário é a de professor. Campello (2003) destaca esta função nos aspectos que compreendem tanto o ensino, como o pensamento crítico. Assim a autora afirma que, neste papel o bibliotecário está:

> "[...] encarregado de ensinar não apenas as habilidades que vinha tradicionalmente ensinando [...] mas está envolvido no desenvolvimento das habilidades de pensar criticamente, [...] ensinando o aprender a aprender."

O que motivou a realização deste estudo foi o frequente questionamento sobre a profissão do bibliotecário em relação ao seu papel no processo de recuperação da informação, visto que o profissional responsável por esta ação está em contato permanente com o usuário.

Promover o uso dos serviços e dos recursos da biblioteca tem sido o foco dos inúmeros treinamentos de usuários nas bibliotecas universitárias. Essa preocupação é destacada por Grogan (1991), quando ele enfatiza a importância de dar autonomia aos usuários e, ao mesmo tempo possibilitar que os treinamentos auxiliem os usuários a se tornarem pessoas confiantes em si mesmas para fazer a busca pela informação desejada. Neste sentido, o autor afirma que "criar uma forma de instrução que seja eficaz enquanto modelo de educação de modo a tornar o usuário autossuficiente, é o grande desafio", para os bibliotecários, sintetiza Grogan (1991, p. 16).

Com isso posto, a questão que motivou a realização deste trabalho compreendeu: qual o papel do bibliotecário na recuperação da informação?

2 OBJETIVOS

O objetivo principal deste estudo é revisar na literatura o papel que o bibliotecário exerce como mediador informacional na educação e no desenvolvimento da competência informacional do usuário no processo de recuperação da informação.

3 PROCEDIMENTOS METODOLÓGICOS

Os procedimentos metodológicos deste estudo teórico compreenderam o levantamento bibliográfico, feito por intermédio de pesquisa exploratória livre acerca do tema em análise, bem como a seleção dos materiais possibilitaram a criação de um referencial teórico sobre a temática. Este levantamento possibilitou às autoras deste trabalho ampliar a compreensão sobre as habilidades do bibliotecário que podem contribuir para o processo de recuperação da informação.

4 RESULTADOS

Destacam-se a seguir os resultados identificados na literatura sobre a temática em discussão.

4.1 Recuperação da informação

Em relação à recuperação da informação recorreu-se a Mooers (1951) e Saracevic (1991). O termo recuperação da informação foi cunhado por Calvin Mooers (1951). Para o autor a recuperação da informação "engloba os aspectos intelectuais da descrição de informações e suas especificidades para a busca além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação". Da mesma forma Saracevic (1991) frisou que, certamente, a recuperação da informação não foi a única responsável pelo desenvolvimento da Ciência da Informação, mas pode ser considerada como o principal fator, ao longo do tempo. Isso significa dizer que, a CI ultrapassou a recuperação da informação, mas os problemas principais tiveram sua origem aí e ainda constituem seu núcleo. Da mesma forma, ressalta-se que uma das grandes preocupações da CI é dar acesso aos usuários das unidades de informação aos recursos de informação.

Observou-se que a recuperação da informação pode ser otimizada com a implementação de atividades de educação de usuários ou competência em informação, temática que será abordada no próximo tópico.

4.2 Competência em informação

Para Takahashi "a sociedade da informação não é um modismo e representa uma profunda mudança na organização da sociedade e da economia, havendo quem a considere um novo paradigma técnico-econômico. É um fenômeno global com elevado potencial transformador das atividades sociais e econômicas, uma vez que a estrutura e a dinâmica dessas atividades inevitavelmente serão, em alguma medida, afetadas pela infraestrutura de informações disponíveis." (2000), p.5)

Neste cenário informacional, pode-se identificar a necessidade de acionar ferramental e técnicas de acessar a informação, de maneira a aperfeiçoar a sua busca e uso. Confirmando esta teoria, Hatschbach (2002, p. 10) afirma que:

> Os indivíduos que pretendem ser agentes de transformação e conquistar seu espaço na sociedade da informação necessitam adquirir habilidades específicas para o trato com a informação no que se refere a sua localização, acesso, uso, comunicação e, principalmente, para a geração de novos conhecimentos.

Percebeu-se na literatura os termos "information literacy", ou competência em informação e "alfabetização informacional" que são tratados a seguir. Neste ponto, surge o tema de estudo que trata especificamente das questões relacionadas à aquisição de competências e conjunto de conhecimentos requeridos na busca, uso e compreensão da informação, voltada para a formação estudantil, profissional e social, denominada information literacy, ou competência em informação.

Antes de abordar especificamente o termo competência em informação apresenta-se a seguir a definição para information literacy. Para isso, recorreu-se à Dudziak (2003). A autora discutiu a definição de information literacy com ênfase no aprendizado ao longo da vida. Para Dudziak (2003, p. 28) Information literacy "é o processo contínuo de internalização de fundamentos conceituais, atitudinais e de habilidades necessárias à compreensão e interação permanente com o universo informacional e sua dinâmica, de modo a proporcionar um aprendizado ao longo da vida".

A expressão information literacy foi utilizada pela primeira vez, de acordo com Dudziak (2003, p. 23), no relatório do bibliotecário americano Paul Zurkowski, intitulado "The information service environmental relation ships and priorities" de 1974. Neste estudo identificou-se o termo que é considerado como a evolução da educação de usuários. Assim, Zurkowski apud Hatschbach (2002, p. 16) afirma que:

> pessoas treinadas para a utilização de fontes de informação em seu trabalho podem ser chamadas de 'competentes em informação' (information literacy). Elas aprendem técnicas e ferramentas informacionais como fontes primárias para encontrarem informação, visando à solução de seus problemas.

O processo de pesquisa, habitualmente exige instrução e o suporte profissional para auxiliar a busca e o uso da informação no ambiente educacional para possibilitar a construção do conhecimento.

A competência informacional é abrangente e a sua implementação demanda o emprego de vários tipos de ferramentas que compreende desde a instrução, o treinamento, a apresentação de interfaces informacionais até a orientação bibliográfica para a realização dos levantamentos necessários para a construção de referenciais teóricos.

De acordo com a American Library Association (1989), os itens essenciais para que a pessoa seja competente em informação são: saber realizar uma busca por informação, fazer a avaliação do resultado, filtrar este resultado e saber usar a informação quando necessário.

Caregnato (2000) define "alfabetização informacional" como atividade, principalmente das bibliotecas acadêmicas, de desenvolver nos alunos as habilidades necessárias para a pesquisa no mundo digital. Da mesma forma, surge destacado o conceito de aprendizagem ao longo da vida, citado por Kuhlthau (2002) apud Hatschbach (2002, p. 18):

> A busca de informação é um processo de construção que envolve a experiência de uma vida, os sentimentos, bem como os pensamentos e as atitudes de uma pessoa.

4.3 O papel do bibliotecário na recuperação da informação

As atividades técnicas de processamento da informação fazem parte do papel do bibliotecário tanto quanto as relacionadas com o processo de transmitir o processo de aprender a aprender, ou seja, compartilhar e multiplicar o que se sabe, visando a formar cidadãos que sabem interagir e usufruir da informação em todos os seus suportes e formas. Essa transferência da informação exige que o bibliotecário se mostre capacitado, aberto a auferir novas habilidades demandadas pelo atual contexto informacional.

Trata-se de um profissional comprometido com a sua vocação, que valoriza a si e ao seu trabalho, segue na educação contínua aliada a novos comportamentos, com novas formas de pensar, que trabalha suas qualidades de criatividade, curiosidade, liderança, empreendedorismo, inovação e sensibilidade para com o público que ele está servindo.

Ao bibliotecário é requerido mais que o domínio dos saberes da biblioteconomia, tendo em vista a complexidade social moderna, a maior necessidade de informação das pessoas, o surgimento de novas tecnologias de informação, a disponibilidade de informação na internet, entre outros fatores.

Com o propósito de auxiliar o bibliotecário a concretizar este amplo espectro de comportamentos profissionais validados por especialistas em programas e projetos para aplicação da competência informacional, que pode resultar em uma recuperação da informação eficaz levou Dudziak (2005) a apresentar um conjunto de elementos que são relevantes e que são indicadas a seguir:

- o bibliotecário deve seu um campeão de causa;
- o centro do processo é o aluno;
- o bibliotecário deve ser um agente educacional;
- é preciso haver cooperação entre docentes e bibliotecários.
- a cultura do livre acesso à informação deve ser enfatizada;
- a inserção no projeto pedagógico;
- definição clara de objetivos e metas;
- planejamento é essencial;
- a transdisciplinaridade e o currículo integrado como marcos para a competência em informação;
- incorporar diferentes espaços de aprendizagem;
- as melhores práticas se constroem no decorrer do processo;
- avaliação constante e controle do processo.

Além disso, destaca-se que o papel do bibliotecário envolve também, desenvolver projetos pedagógicos na área da competência em informação, com o objetivo de educar os usuários para a busca, recuperação e uso da informação, fazendo o papel de um agente educacional em trabalho conjunto com os professores da instituição.

5 CONSIDERAÇÕES FINAIS

Percebeu-se neste estudo, que uma importante mudança se passa no perfil do profissional bibliotecário, ele deve estar atento e buscar inserção nesse novo cenário. E o profissional moderno está diante de uma ótima oportunidade, para, munido dos requisitos necessários, ampliar a visão da profissão e se lançar aos desafios da competência informacional.

Os mediadores da informação, no caso os bibliotecários, e o papel desempenhado por eles junto aos usuários, em vista da recuperação da informação, recebem enorme valorização na atualidade, marcada por informação em excesso, emitida por canais diversos e crescentemente modernos. Essa lacuna é ocupada pela habilidade do profissional em filtrar o que realmente é útil aos usuários e também capacitá-los a serem bem sucedidos no universo informacional moderno.

Na sociedade da informação, o tema educação de usuários/competência informacional e a recuperação da informação, é visto como essencial. Um grande contingente de informação é ofertado aos usuários, porém, geralmente ele não está apto a realizar busca e recuperação que atenda às suas reais necessidades.

Verificado este fato, este trabalho mostrou como deve ser aplicada a educação de usuários na atualidade, com vistas à recuperação da informação. Constatou-se que o desenvolvimento da competência informacional, que é a educação de usuários expandida, é de suma importância, pois vai além da biblioteca, viabilizando o aprender por toda a vida.

O papel do bibliotecário neste processo também foi abordado. Apresentou a importância capital do profissional da informação no desenvolvimento da habilidade informacional dos usuários e as novas competências requeridas atualmente para um excelente trabalho junto aos mesmos.

Desta forma, o trabalho respondeu às questões propostas, por meio da exposição da relevância da obtenção da competência informacional e o papel do bibliotecário que, de fato, com os seus saberes, contribui para aprimorar o processo de recuperação da informação.

REFERÊNCIAS

AMARAL, S. A. do. *Marketing*: abordagem em unidades de informação. Brasília: Thesaurus, 1998, 224 p.

BELLUZZO, R. C. B., MACEDO, N. D. de. A gestão da qualidade em serviços de informação: contribuição para uma base teórica. *Ciência da Informação*, v. 22, n. 2, p. 124-132, maio/ago., 1993.

BELLUZZO, R. C. B. Contribuição da competência em informação em bibliotecas públicas paulistas: uma experiência com apoio de oficinas de trabalho. In: *CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO*, 21, jul. 2005, Curitiba: FEBAB, 2005. 1 CD-ROM.

CAMPELLO, B. A competência informacional na educação para o século XXI. In: *biblioteca escolar:* temas para a prática pedagógica. Belo Horizonte: Autêntica, 2002. p. 9-11.

CAMPELLO, B. O movimento da competência informacional: uma perspectiva para o letramento informacional. *Ciência da Informação*, Brasília, v. 32, n. 3, p. 28-37, set./dez. 2003.

CAREGNATO, S. E. O desenvolvimento de habilidades informacionais: o papel das bibliotecas universitárias no contexto da informação digital em rede. *Revista de Biblioteconomia & comunicação*, Porto Alegre, v. 8, p. 47-55, 2000.

DUDZIAK, E. A. Information literacy: princípios, filosofia e prática. *Ciência da Informação,* Brasília, v. 32, n. 1, p. 23-35, 2003.

DUDZIAK, E. A.; GABRIEL, M. A.; VILLELA, M. C. A educação de usuários de bibliotecas universitárias frente à sociedade do conhecimento e sua inserção nos novos paradigmas educacionais. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 11, 2000, Florianópolis. *Anais*.

Florianópolis: UFSC, 2000. Disponível em: <http://snbu.bvs.br/snbu2000/parallel.html>.Acesso em 28 set. 2018.

GROGAN, D. *A prática do Serviço de Referência.* Brasília. Briquet de Lemos, 1995. 196 p.

HATSCHBACH, M. H. DE L. *Information literacy:* aspectos conceituais e iniciativas em ambiente digital para o estudante de nível superior. Rio de Janeiro, 2002. 108 f. Dissertação (Mestrado em Ciência da Informação) – UFRJ/ECO-MCT/IBICT, Rio de Janeiro, 2002.

KUHLTHAU, C. *Seeking meaning.* Norwood: Ablex, 1993.

KUHLTHAU, C. *Como usar a biblioteca na escola:* um programa de atividades para o ensino fundamental. 2. ed. Belo Horizonte: Autêntica, 2004.

LANCASTER, F. W. Avaliação da instrução bibliográfica. In: _____. *Avaliação de serviços de bibliotecas.* Brasília: Briquet de Lemos Livros, 1996. Cap. 12, p. 226-262.

MÜELLER, S. P. M. Perfil do bibliotecário, serviços e responsabilidades na área de informação e formação profissional. *Revista de Biblioteconomia de Brasília,* v. 17, n. 1, p. 63-70, jan./jul. 1989.

PASQUARELLI, M. L. R. *Procedimentos para busca e uso da informação:* capacitação do aluno de graduação. Brasília: Thesaurus, 1996. 86 p.

PASQUARELLI, M. L. R. *O papel da universidade na capacitação do estudante de graduação para busca e uso da informação*: a disciplina Orientação Bibliográfica em revisão. 1993. 115f. Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicação e Artes. Universidade de São Paulo, São Paulo, 1993.

SADI, B. S. C. *et al*. Satisfação e frustração do usuário em obter documentos em uma biblioteca acadêmica de saúde pública. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS 4, 1985, Campinas. *Anais*. Campinas, 1985.

SANTOS, J. P. O moderno profissional da informação: o bibliotecário e seu perfil face aos novos tempos. *Informação & Informação,* Londrina, v. 1, n. 1, p. 5-13, jan./jun. 1996.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. *Perspect. ciênc. inf.*, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235/22>. Acesso em: 28 set. 2018.

TAKAHASHI, T. (Org.) *Sociedade da informação no Brasil:* livro verde. Brasília: Ministério da Ciência e Tecnologia, 2000. 203 p.

TARAPANOFF, K. *Perfil do profissional da informação no Brasil*: diagnóstico de necessidade de treinamento e educação continuada. Brasília: IEL/DF, 1997. 137 p.

TARGINO, M. G. Quem é o profissional da informação? *Transinformação.* Campinas, v. 12, n. 2, p. 61-69, jul./dez. 2000.

O PAPEL DOS METADADOS NA CURADORIA DIGITAL

THE ROLE OF METADATA IN DIGITAL CURATION

Rachel Cristina Vesu Alves[1]
(1) Faculdade de Filosofia e Ciências - UNESP - Universidade Estadual Paulista, Av. Hygino Muzzi Filho, 737, Bairro: Mirante, CEP: 17.525-900 - Marília, SP, rachel.vesu@unesp.br

Resumo:
Os metadados são importantes para proporcionar a organização, a representação, o acesso, a busca, a recuperação e a preservação dos recursos informacionais nos diferentes domínios do conhecimento e ambientes informacionais digitais. Os metadados podem desempenhar diversas funções, sendo que uma das mais requeridas nos ambientes informacionais na atualidade refere-se a preservação digital e curadoria digital. Assim, o problema norteador deste trabalho está em identificar qual o papel dos metadados na curadoria digital? Por meio de uma metodologia exploratória, o objetivo deste trabalho é abordar a importância e o papel dos metadados na curadoria digital, assim como, identificar quais os principais tipos ou categorias de metadados mais utilizados na curadoria digital. Os resultados parciais deste trabalho, revelam uma variedade de papéis ou funções que os metadados podem exercer no cenário da curadoria digital, assim como, uma variedade de padrões a serem utilizados, dependendo do contexto ou domínio. Deste modo, a próxima fase do estudo encaminha-se para a identificação dos padrões de metadados que podem ser utilizados no cenário de curadoria digital.
Palavras-chave: Metadados. Curadoria Digital. Preservação Digital.

Abstract:
Metadata is important to provide the organization, representation, access, search, retrieval and preservation of information resources in different domains of knowledge and digital information environments. Metadata can perform several functions, and one of most required in today's information environments refers to digital preservation and digital curation. Thus, the guiding problem of this work lies in identifying the role of metadata in digital curation. Through an exploratory methodology, the objective of this research is to approach the importance and role of metadata in digital curation, also identify main types or categories of metadata most used in digital curation. The partial results of this research reveal a variety of roles or functions that metadata can exert in digital curatorial scenario, as well, as a variety of standards to be used depending on context or domain. Thus, next phase of the study is directed to identify metadata standards that can be used in digital curation scenario.
Keywords: Metadata. Digital Curation. Digital Preservation.

1 INTRODUÇÃO

É indiscutível a importância que os metadados adquiriram na atualidade para proporcionar a organização, a representação, o acesso, a busca, a recuperação e a preservação nos diferentes domínios do conhecimento e ambientes informacionais digitais.

Os metadados podem desempenhar diversas funções como, por exemplo, identificação e descrição da informação; busca e recuperação; localização, restrição de uso e formas de acesso; autoria e propriedade intelectual; visibilidade e acessibilidade das informações; valoração do conteúdo, entre outras. Dentre as funções desempenhadas pelos metadados, a que vem sendo mais requerida atualmente nos ambientes digitais refere-se a preservação digital (MÉNDEZ RODRÍGUEZ, 2002).

A preservação digital pode ser definida como uma "[...] série de atividades gerenciais necessárias para garantir o acesso contínuo aos materiais digitais ao longo do tempo [...] refere-se a todas as ações necessárias para manter o acesso aos materiais digitais, além dos limites de descontinuidade de mídia ou mudança tecnológica e organizacional." (BEAGRIE; KILBRIDE, 2015, tradução nossa).

A preservação digital pode ser considerada como um aspecto importante da curadoria digital, outro termo bastante difundido nos últimos tempos. A curadoria digital está relacionada com o gerenciamento de dados de pesquisa, em outras palavras refere-se a "[...] manter, preservar e agregar valor aos dados de pesquisa digital em todo o seu ciclo de vida." (DIGITAL CURATION CENTER, 2018).

Houve um crescimento nos últimos tempos não só dos recursos informacionais digitais, mas também de uma grande quantidade de dados que precisam ser gerenciados para serem acessados ao longo do tempo em um ambiente digital.

A curadoria digital consiste em um conceito mais amplo, pois "[...] começa antes que os objetos digitais sejam criados, configurando o planejamento da coleta de dados que resulta em objetos digitais "preparados para curadoria", que estão na melhor condição possível para garantir que possam ser mantidos e utilizados no futuro."

De acordo com Harvey (2010) a curadoria digital não engloba apenas a questão da preservação, mas também o reuso dos dados. Em outras palavras, "A curadoria digital enfatiza a adição de valor a conjuntos de dados e objetos digitais, por meio de itens como metadados ou anotações adicionais, para que possam ser reutilizados." (HARVEY, 2010, p. 08).

Nesse sentido, pode-se entender que o uso de metadados passa a ser essencial para a curadoria digital, pois exige um gerenciamento dos recursos em seu ciclo de vida; deve proporcionar o acesso aos dados e recursos informacionais; a preservação digital dos dados e recursos informacionais ao longo do tempo, além do reuso dos dados.

Assim, o problema norteador deste trabalho está em identificar qual o papel dos metadados na curadoria digital?

2 OBJETIVOS

O objetivo deste trabalho é abordar a importância e o papel dos metadados na curadoria digital, assim como, identificar quais os principais tipos ou categorias de metadados mais utilizados na curadoria digital.

3 PROCEDIMENTOS METODOLÓGICOS

A metodologia da pesquisa está embasada em estudo exploratório que, segundo Cervo e Bervian (2003), consiste na busca de informações sobre um determinado assunto, para que solucione os diversos aspectos de um problema. Segundo Köcher (2002, p. 126) estudos exploratórios desencadeiam "[...] um processo de investigação que identifique a natureza do fenômeno e aponte as características essenciais das variáveis que se quer estudar".

Assim, acredita-se ser possível identificar a importância e o papel dos metadados na curadoria digital.

4 RESULTADOS

Os resultados parciais revelam uma variedade de papéis ou funções que os metadados podem exercer no cenário da curadoria digital. Assim como, uma diversidade de padrões de metadados a serem utilizados, dependendo do contexto e domínio no qual está inserido.

A preservação é uma das funções que os metadados desempenham na curadoria digital. Segundo Grácio (2012) a preservação digital pode ser abordada sob três aspectos principais: o aspecto organizacional, o aspecto legal e o aspecto técnico. Os metadados permeiam todos esses aspectos, pois devem ser determinados de modo amplo na política de preservação digital das instituições, que por sua vez, engloba a determinação desses três aspectos.

Porém, são nos aspectos técnicos da preservação onde estão presentes a maior parte dos metadados responsáveis pela descrição e preservação digital do recurso informacional.

A obsolescência tecnológica de software e hardware, dos formatos de arquivo, além da degradação das mídias digitais, são fatores preponderantes que motivam o estabelecimento da preservação digital e do acesso a longo prazo dos recursos informacionais em ambientes digitais.

Assim, torna-se necessário o uso de estratégias de preservação (migração, emulação, encapsulamento etc) para preservar o aspecto físico, de conteúdo e a autenticidade dos recursos informacionais, garantindo seu acesso ao longo do tempo.

Neste contexto, os metadados e padrões de metadados a serem utilizados na preservação digital estão mais relacionados com as funções de: descrição, organização, mudança de formatos de arquivo, informações sobre as estratégias de preservação utilizadas e as ações de preservação sobre o recurso informacional ao longo do tempo.

É importante ressaltar também que algumas informações a serem preservadas ao longo do tempo estão representadas em metadados descritivos. Um exemplo a ser considerado são os metadados título, autor e data, geralmente presentes em padrões de metadados descritivos. Atributos descritivos, essenciais para caracterizar um recurso informacional, são dados a serem preservados ao longo do tempo, além dos atributos próprios da preservação digital como, por exemplo, atributos sobre as estratégias de preservação utilizadas.

De acordo com o *Digital Curation Center* (2018 tradução nossa) o ciclo de vida da curadoria digital compreende as seguintes etapas:

- **Conceitualizar:** conceber e planejar a criação de objetos digitais, incluindo métodos de captura de dados e opções de armazenamento;
- **Criação:** produção de objetos digitais e atribuição de metadados administrativos, descritivos, estruturais e técnicos para arquivamento;
- **Acesso e uso:** garantir que os usuários designados possam acessar facilmente os objetos digitais [...];
- **Avaliar e selecionar:** avaliar objetos digitais e selecionar aqueles que requerem curadoria e preservação a longo prazo. Seguir as orientações documentadas nas políticas e requisitos legais;
- **Descartar:** liberar os sistemas de objetos digitais não selecionados para curadoria e preservação a longo prazo [...];
- **Transferir:** objetos digitais para um arquivo, repositório digital confiável, data center ou similar, seguindo novamente as diretrizes documentadas, políticas e requisitos legais;
- **Ação de preservação:** empreender ações para assegurar a preservação e retenção a longo prazo da natureza autêntica dos objetos digitais;
- **Reavaliar:** objetos digitais que falham nos procedimentos de validação para avaliação [...];
- **Armazenar:** manter os dados de maneira segura, conforme descrito pelos padrões relevantes;
- **Acesso e reutilização:** garantir que os dados sejam acessíveis aos usuários designados para uso e reutilização [...];
- **Converter:** criar novos objetos digitais a partir do original, por exemplo, migrando para uma forma diferente.

Estão contempladas em todas essas etapas vários tipos de metadados. Assim, observa-se que os metadados vão desempenhar papéis importantes no ciclo de vida da curadoria digital.

De acordo com Harvey (2010, p. 67-68, tradução nossa), pode-se encontrar os seguintes tipos de metadados na curadoria digital:

- **Metadados administrativos:** metadados relacionados ao uso, gerenciamento e processo de codificação de objetos digitais durante um período de tempo. Inclui os subconjuntos de metadados técnicos, metadados de gerenciamento de direitos e metadados de preservação.
- **Metadados descritivos:** metadados que descrevem uma obra para fins de descoberta e identificação, como criador, título e assunto.
- **Metadados técnicos:** uma forma de metadados administrativos que lida com o processo de codificação de criação ou armazenamento de formatos do recurso.
- **Metadados estruturais:** metadados que indicam como os objetos compostos são estruturados, provê o suporte para o uso dos objetos.
- **Metadados de preservação:** metadados administrativos que lidam com a proveniência de um recurso e sua gestão de arquivamento. (HARVEY, 2010, p. 67-68, tradução nossa).

Os tipos e categorias de metadados citados estarão presentes nos esquemas dos padrões de metadados. Alguns padrões serão compostos por mais de um tipo de metadados, já outros apresentarão a predominância de um desses tipos. É o caso do padrão de metadados de preservação digital PREMIS (*Data Dictionary for Preservation Metadata*).

4 CONSIDERAÇÕES FINAIS

Este trabalho consiste em uma análise preliminar do papel dos metadados na curadoria digital.

Foi possível identificar que os metadados estão presentes em todo ciclo de vida da curadoria digital e que se constituem em diferentes tipos como, administrativos, descritivos, técnicos, estruturais e de preservação.

A preservação digital constitui-se como um dos aspectos da curadoria e, mesmo voltada para aspectos mais técnicos dos recursos informacionais e requisitos do sistema, também preservam informações descritivas e de contextualização de um recurso, por exemplo.

Deste modo, a próxima fase do estudo encaminha-se para a identificação dos padrões de metadados que podem ser utilizados no cenário de curadoria digital, considerando a variedade de domínio e padrões existentes. Assim, faz-se necessário a identificação desses padrões com base nas características específicas de domínio e as diretrizes estabelecidas pela *Digital Curation Center*.

REFERÊNCIAS

BEAGRIE, N.; KILBRIDE, M. *Digital Preservation Handbook*. Disponível em: <https://dpconline.org/handbook >. Acesso em: 07 out. 2018.

CERVO, A. L.; BERVIAN, P. A. Metodologia científica. 5. ed. São Paulo: Prentice Hall, 2003.

DIGITAL CURATION CENTER. *What is digital curation?* 2018. Disponível em: <http://www.dcc.ac.uk/digital-curation/what-digital-curation>. Acesso em: 07 out. 2018.

GRÁCIO, J. C. A. Preservação digital na gestão da informação: um modelo processual para as instituições de ensino superior. São Paulo: Cultura Acadêmica, 2012. ISBN 9788579833335. Disponível em: <http://hdl.handle.net/11449/113727>. Acesso em: 22 nov. 2018.

HARVEY, R. *Digital Curation:* a how-to-do-it maual. New York: Neal-Schuman, 2010.

KÖCHE, J. C. Fundamentos da metodologia Científica: teoria da ciência e iniciação à pesquisa. 20.ed. Petrópolis: Vozes, 2002.

MÉNDEZ RODRÍGUEZ, E. *Metadados y recuperación de información*: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: Trea, 2002. 429 p.

O TERMO *BIG DATA*:
quebra de paradigma dos n-V's

THE TERM BIG DATA:
paradigm break of n-V's

Marcos de Souza[1], Fernanda Gomes Almeida[2], Renato Rocha Souza[3].
(1) Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, marcosdesouza82@gmail.com.
(2) Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, usernanda@gmail.com.
(3) Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627 / Fundação Getúlio Vargas, Praia de Botafogo, 190 - Botafogo, Rio de Janeiro – RJ, rsouza.fgv@gmail.com.

Resumo:
Os dados são considerados o novo petróleo da era digital. O crescimento e integração de grandes volumes de dados digitais – *Big Data* - têm sido utilizados para tomada de decisão em diferentes áreas. Entretanto, percebe-se que não existe uma universalização do conceito de *Big Data*, sobretudo, no meio acadêmico por se tratar de um termo relativamente novo. A pesquisa tem como objetivo identificar os conceitos relacionados ao termo *Big Data* e; identificar as características que compõem o conceito. A pesquisa foi elaborada com base em artigos científicos e livros digitais e impressos publicados entre os anos de 2011 e 2018. Como resultado percebe-se que a literatura apresenta tradicionalmente 5V's como características de um *Big Data*, sendo volume, variedade, velocidade, veracidade e viabilidade. Contudo, a pesquisa permitiu identificar mais 2V's e diversos conceitos para o termo, o que não permite assim uma consolidação do termo. Com o avanço tecnológico, novas possibilidades têm surgido e colocado as características em cheque. Sugere-se como pesquisas futuras uma caracterização real do termo de forma que o mesmo não esteja atrelado a determinada letra alfabética.
Palavras-chave: *Big Data*. Conceito. Viabilidade. Venalidade. n-V's.

Abstract:
Data is considered the new oil of the digital era. The growth and integration of large volumes of digital data - Big Data - have been used for decision making in several areas. However, this is not an international model of big data. A research aims to identify the data related to the term Big Data and; identify as characteristics that make up the concept. The research was based on scientific and printed data between the years 2011 and 2018. As a result it realizes that a literature is traditionally produced as a large volume of data, being volume, variety, speed, veracity and feasibility. In this article, you may have more information about the concepts of a term that is not mandatory. With technological technology, new features have emerged and placed as features in check. It is suggested as a new form of characterization of the term so that it is not tied to an alphabetic letter.
Keywords: Big data. Concept. Viability. Venality. n-V's.

1 INTRODUÇÃO

Os dados são considerados o novo petróleo na era digital, possuindo um valor altamente estratégico. Tão importante quanto o petróleo, os dados são a arma mais poderosa da sociedade. Grandes empresas tecnológicas como *Amazon*, *Apple*, *Facebook*, *Google* e *Microsoft* apresentaram um faturamento maior que 25 bilhões de dólares apenas no primeiro semestre de 2017, demonstrando que as organizações que detêm os dados também detêm o poder (ECONOMIST, 2017).

> A produção e armazenamento de dados têm passado por um processo evolutivo desde a popularização dos computadores e a difusão da internet, dessa forma, tem ganhando destaque a partir de uma produção massiva de dados realizada por meio de sites, redes sociais digitais, *smartphones*, sistemas informatizados, satélites, sensores de carros, aviões e trens, câmeras de segurança entre outros tipos de serviços como *streaming* (AMARAL, 2016). Além da evolução referente a capacidade de armazenamento de dados, faz necessário destacar também o quantitativo de vezes que os dados são compartilhados pelos usuários (MARQUESONE, 2016).

O conceito de *Big Data foi cunhado a partir do* crescimento e integração de grandes volumes de dados digitais produzidos por diversos meios e tecnologias, bem como pelo uso de ferramentas que perpassam as etapas de produção, compartilhamento, curadoria e análise de maneira eficiente dessa gigantesca quantidade de dados, resultando em informações de valia e permitindo descobrir novos padrões significativos de conhecimentos (SAKR, 2016; ELSHAWI; SAKR, 2018).

Buscas realizadas utilizando o termo *Big Data* nas ferramentas de pesquisa acadêmica *Google Scholar* e o Portal de Periódicos da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) apresentou como resultado, respectivamente, 4.820.000 e 1.739.128 documentos que abordam de alguma forma a temática *Big Data*. Já a pesquisa realizada no buscador *Google* apresentou aproximadamente 4.170.000.000 resultados. Assim, trata-se de um quantitativo bastante representativo, uma vez que se refere a um termo que foi cunhado há poucos anos.

A definição de *Big Data* é considerada pobre (BOYD; CRAWFORD, 2012), equivocada e genérica, pois refere-se apenas ao tamanho (DAVENPORT, 2017). Contudo, este trabalho parte do princípio que o *Big Data* é um termo em fase de amadurecimento e que ainda está em processo de consolidação de conceitos.

Justifica-se a importância desse trabalho pela notoriedade do termo *Big Data* nos últimos anos, tanto no meio acadêmico quanto empresarial e pela discrepância entre os conceitos encontrados na literatura. Como hipótese, acredita-se que, por se tratar de um termo relativamente novo e principalmente por estar atrelado aos meios e avanços tecnológicos, possam surgir cotidianamente novas características que contribuem para essa diversidade de conceitos.

2 OBJETIVOS

O trabalho tem como objetivo identificar os conceitos relacionados ao termo *Big Data* e, identificar as características que compõem o conceito.

3 PROCEDIMENTOS METODOLÓGICOS

Esta é uma pesquisa de natureza básica, abordando o problema de forma qualitativa. Do ponto de vista dos seus objetivos, é classificada como exploratória e descritiva. Da perspectiva dos procedimentos técnicos, identifica-se como pesquisa bibliográfica (GIL, 2010; Gil, 2016).

A pesquisa consiste em um levantamento, análise e seleção de trabalhos que trazem uma conceituação do termo *Big Data*, bem como suas características, publicados entre os anos de 2011 e 2018.

Os trabalhos foram identificados a partir de pesquisas realizadas no Portal de Periódicos da CAPES e no *Google Scholar*. Além desses, foram utilizados artigos científicos, livros digitais e impressos da área de domínio.

As publicações selecionadas serviram como base para a identificação dos conceitos relacionados ao termo *Big Data*, bem como a identificação das características que compõem o seu conceito.

4 RESULTADOS

Um número expressivo de documentos acadêmicos apresenta uma série de conceitos sobre o termo *Big Data*. Esses conceitos são apresentados no Apêndice A. Verifica-se dessa forma, uma diversidade de conceitos o que pode contribuir para a não consolidação do termo.

A consultoria Gartner desenvolveu um estudo no ano de 2001 no qual apresentou o volume, a variedade e a velocidade - conhecidos como 3V's - como um dos maiores desafios para a manipulação e a gestão de dados. Esse viés tem sido discutido com maior relevância no meio acadêmico (Laney, 2001). Posteriormente, a veracidade (qualidade) valor (relevância) foram incorporadas aos 3V's e constituíram assim o 5V's. Essas características têm ganhado destaque nas investigações relacionadas ao meio empresarial (Assunção et al., 2015). Tais características são utilizadas para definir o termo *Big Data*.

Contudo, além dos 5V's, outras características podem surgir, formando assim os n-V's, uma vez que o conceito de *Big Data* está atrelado aos avanços tecnológicos.

4.1 Big Data e suas características

O *Big Data* é um fenômeno tecnológico e humano no qual são produzidos dados em diversos formatos e que são armazenados em grandes quantidades de dispositivos de armazenamento. Refere-se às características inerentes de um *Big Data* os n-V's, sendo eles:

- Volume - relaciona-se ao quantitativo de informações armazenadas em dispositivos físicos ou virtualizados. No atual cenário, os armazenamentos de dados perpassam por um crescimento exponencial mediante a tecnologia de computação nas nuvens. Os dados têm sido criados em escala de *zetabytes* por meio de diversas fontes e aplicações por atividades cotidianas (VASCONCELOS;

BARÃO, 2017). Considerando que bilhões de pessoas geram informações diariamente por meio da internet, estima-se que o volume de informações produzidas possa dobrar a cada 18 meses. O atributo volume é considerado a característica mais importante no conceito de *Big Data* e faz dimensão sem precedentes do volume de dados (MACHADO, 2018).

- Variedade - define a diversidade de informações e dados produzidos e que possuem formatos heterogênicos, podendo ser dados estruturados, não estruturados ou semiestruturados que são criados, compartilhados e consumidos em diferentes formatos e uma variedade de fontes de dados, como por exemplo *upload* de fotos, transmissão de vídeos em tempo real, textos publicados em redes sociais (MACHADO, 2018). A variedade também indica a variação das áreas de aplicações de um *Big Data*, tais como as áreas governamentais, financeiras, transportes e automação, varejistas, seguros e outras possibilidades (VASCONCELOS; BARÃO, 2017);
- Velocidade - além dos desafios referentes ao volume e variedade de dados, o *Big Data* faz referência a agilidade em que os dados são coletados, processados, analisados e utilizados (MARQUESONE, 2016; VASCONCELOS; BARÃO, 2017; MACHADO, 2018). Um fator que contribuiu positivamente para a característica da velocidade está na evolução das telecomunicações, tais como largura de banda e infraestrutura física, possibilitando assim que os dados sejam transportados com uma maior velocidade por meio da internet (VASCONCELOS; BARÃO, 2017);
- Veracidade - diz respeito a qualidade dos dados e a confiabilidade sobre as fontes dos dados. O *Big Data* está inserido em um contexto de variedade e volume de dados, sendo comum a existência de dados inconsistentes. Dessa forma, torna-se necessário garantir a autenticidade das fontes e dos dados coletados de forma que sejam confiáveis para uma determinada solução (MARQUESONE, 2016; VASCONCELOS; BARÃO, 2017; MACHADO, 2018);
- Valor - caracteriza-se pelo quão valioso e significativo um dado pode ser para uma determinada solução em um *Big Data*, obtendo assim valor qualitativo a partir do volume, variedade e velocidade de dados (MARQUESONE, 2016). Torna-se importante ressaltar que se trata da relevância dos dados após o processamento. Os dados brutos perpassam por análise e processamento onde poderão ser acrescidos de valor, transformando-se em vantagem competitiva para negócios. (VASCONCELOS; BARÃO, 2017; MACHADO, 2018).
- Viabilidade - destina-se a identificar o relacionamento entre variáveis e padrões latentes de dados (BIEHN, 2013). Esse conceito pode ser discutido, por exemplo, em modelagem de tópicos que faz uso de *machine learning* e processamento de linguagem natural em grandes corpora textuais para identificar a emersão dos tópicos latentes.
- Venalidade - projetado futuramente em formatos que possam ser vendidos (DAVENPORT, 2017). A questão da venda de dados já ocorre por meio de empresas prestadoras de serviços *online*, entretanto, o autor discute a questão de um formato específico para o conceito de grandes volumes de dados.

5 CONSIDERAÇÕES FINAIS

O termo *Big Data*, novo se comparado a outros conceitos científicos consolidados na literatura, apresenta uma diversidade de conceitos, contribuindo para a não consolidação do termo.

Esse fato pode ser observado no Apêndice A que mostra uma não uniformidade entre as definições elaboradas pelos diversos autores, inclusive com críticas, disparidades, simplicidade e homogeneidade.

Outro fator que contribui para falta de consolidação do termo *Big Data* está nas suas características. Inicialmente conhecidas por 3V's - volume, variedade e velocidade; mais tarde com o acréscimo de veracidade e viabilidade tornando-se 5V's; porém, as características que compõem o conceito de *Big Data* ultrapassam esse quantitativo, podendo ser acrescentado também viabilidade e venalidade, tornando-se assim 7V's.

Essa evolução de características de palavras iniciadas com a letra "V" tem acontecido mediante ao surgimento de novas necessidades que aparecem ao se realizar algum tipo de processamento com *Big Data* e também por conta do avanço tecnológico. Entretanto, acredita-se que as tecnologias estejam longe de serem estabilizadas. Com isso, novas possibilidades ou novos "V's" poderão surgir nos próximos anos, como por exemplo, a Variabilidade – capacidade de submeter-se a variações de mudanças; e Visualização de dados associados ao *Big Data*.

Percebe-se que as características associadas ao conceito de um *Big Data* no que diz respeito aos 5V's tem se apresentado de forma sólida na literatura, entretanto, com o avanço tecnológico, novas possibilidades têm surgido e colocado as características em cheque. Dessa maneira, surge o que podemos chamar de n-V's, um paradigma ainda em construção, aberto para as novas características ou representações do termo *Big Dat*a.

Ressalta-se que, em busca realizada no Portal de Periódicos da CAPES utilizando os termos "n-V's Big Data", "nV's Big Data", "n-Vs Big Data" e "nVs Big Data", apesar da quantidade elevada resultados que tratam do Big Data, não foram encontrados resultados que abordem os n-V's da forma que foi proposta neste trabalho – o paradigma dos n-V's.

Sugere-se como pesquisas futuras uma quebra de paradigmas com relação à caracterização real do termo de forma que o mesmo não esteja atrelado à determinada letra alfabética, e sim ao que de fato representa um *Big Data.* Além disso, perspectivas futuras sobre o *Big Data* e tecnologias podem contribuir para a formação de um conceito consolidado, embora acredita-se que possa levar anos para realização de uma definição universalizada do conceito.

REFERÊNCIAS

AMARAL, F. *Introdução à ciência de dados*: mineração de dados e big data. Rio de Janeiro: Alta Books, 2016.

ASSUNÇÃO, M. D. *et al*. Big Data computing and clouds: trends and future directions. *Journal of Parallel and Distributed Computing*, v. 79, p. 3-15, 2015.

BIEHN, N. The missing V's in big data: viability and value. *Wired,* New York, 2013. Disponível em: <https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value>. Acesso em: 10 set. 2018.

BOYD, D.; CRAWFORD, K. Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, v. 15, n. 5, p. 663, 2012.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR. *Portal de Periódicos*. Disponível em: <http://www.periodicos.capes.gov.br/>. Acesso em: 05 set. 2018.

CRAWFORD, K.; SCHULTZ, J. Big Data and Due Process: toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review*, v. 55, n. 1, p. 93-128, 2014. Disponível em: <http://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4>. Acesso em: 11 set. 2018.

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

ECONOMIST. *The world's most valuable resource*: data and the new rules of competition. London, v. 423, n. 9039, 2017.

ELSHAWI, R.; SAKR, S. Big Data Systems Meet Machine Learning Challenges: towards Big Data Science as a Service. *Big Data Research*, v. 5, n. 10, p. 7, 2018. Disponível em: <https://arxiv.org/pdf/1709.07493.pdf>. Acesso em: 5 set. 2018.

FRICKÉ, M. Big data and its epistemology. *Journal of the Association for Information Science and Technology*, v. 66, n. 4, p. 651-661, 2015.

GANTZ, J.; REINSEL, D. Extracting value from chaos. *IDC Review*, v. 1142, n. 2011, p. 1-12, 2011. Disponível em: <https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos--ar.pdf>. Acesso em: 09 set. 2018.

BIG DATA. In: IT GLOSSARY. Stamford: Gartner Group, 2012. Disponível em: <https://www.gartner.com/it-glossary/big-data/>. Acesso em: Acesso em 09 set. 2018.

GEORGE, G.; HAAS, M.R.; PENTLAND, A. Big Data and Management. *Academy of Management Journal*, 2014, Vol. 57, No. 2, 321–326. Disponível em: <http://dx.doi.org/10.5465/amj.2014.4002>. Acesso em: 07 set. 2018.

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

____. *Métodos e técnicas de Pesquisa Social*. 6. ed. São Paulo: Atlas, 2016.

GOMES, R.D.P. *Big Data*: desafios à tutela da pessoa humana na sociedade da informação. Rio de Janeiro: Lumen Juris, 2017.

GOOGLE. Disponível em: <http://www.google.com/>. Acesso em: 05 set. 2018.

GOOGLE SCHOLAR. Disponível em: <https://scholar.google.com.br/>. Acesso em: 05 set. 2018.

HIJMANS, H. *The european union as guardian of internet privacy*. Brussels: Springer International Publishing, 2016.

International Business Machines (IBM). O que é Big Data, 2014. Disponível em: <https://i.pinimg.com/originals/f6/5c/fd/f65cfdfe4813b82155e9a35eacf590f4.jpg>. Acesso em: 10 set. 2018.

INSTITUTO DE TECNOLOGIA & SOCIEDADE DO RIO (ITSRio). Big Data no projeto Sul Global: relatório sobre estudos de caso. Rio de Janeiro: p.26, 2016. Disponível em: <https://itsrio.org/wp-content/uploads/2016/03/ITS_Relatorio_Big-Data_PT-BR_v2.pdf>. Acesso em: 9 set. 2018.

LANEY, D. 3D data management: controlling data volume, velocity and variety. *META Delta*, Stamford, 6 fev. 2001. Application Delivery Strategies. Disponível em: <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Acesso em 9 set. 2018.

MACHADO, F.N.R. *Big Data*: o futuro dos dados e aplicações. São Paulo: Érica, 2018.

MAYER-SCHÖNBERGER, V.; CUKIER, K. *Big Data*: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro: Elsevier, 2013.

MARQUESONE, R. *Big Data*: técnicas e tecnologias para extração de valor dos dados. São Paulo: Casa do Código, 2016.

PROVOST, F.; FAWCETT, T. *Data Science para negócios*: o que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro: Alta Books, 2016.

SAKR, S. *Big data 2.0 processing systems:* a survey. Springer International Publishing, 2016.

TAURION, C. *Big data*. Brasport, 2013.

VASCONCELOS, J. B.; BARÃO, A. *Ciência dos dados nas organizações*: aplicações em python. Lisboa: FCA, 2017.

VOLPATO, T; RUFINO, R.R.; DIAS, J.W. *Big data - transformando dados em decisões*. 2014. Disponível em: <http://web.unipar.br/~seinpar/2014/artigos/graduacao/Tiago_Volpato.pdf>; Acesso em: 9 set. 2018.

Apêndice A – Conceitos de ***Big Data***

Referência	Conceitos
(GANTZ; REINSEL, 2011)	"[...] as tecnologias de Big Data descrevem uma nova geração de tecnologias e arquiteturas projetadas para extrair economicamente o valor de volumes muito grandes e uma variedade de dados, permitindo alta velocidade de captura, descoberta, e/ou análise."
(BOYD; CRAWFORD, 2012)	"Big Data é, em vários sentidos, um termo pobre." "Definimos big data como um fenômeno cultural, tecnológico e acadêmico." "Big Data é menos sobre dados que são grandes do que é sobre a capacidade de pesquisa, agregação e referência cruzada de grandes conjuntos de dados."
(GARTNER GROUP, 2012)	"Big Data são ativos de informações de grande volume, alta velocidade e / ou de alta variedade que exigem formas inovadoras e econômicas de processamento de informações que permitem uma melhor percepção, tomada de decisões e automação de processos."
(EUROPEAN COMMSION, 2013)	"Big Data se refere ao crescimento exponencial tanto da disponibilidade quanto no uso automatizado de informação: refere-se a conjuntos de dados digitais gigantescos detidos por empresas, governos e outras organizações de grande porte, que são amplamente analisados (daí o nome: *analytics*) usando algoritmos de computador. Big Data pode ser usado para identificar tendências mais gerais e correlações, mas também pode ser processado, de modo a afetar diretamente os indivíduos."
(MAYER-SCHÖNBERGER; CUKIER, 2013)	"Big Data refere-se a coisas que se podem fazer em grande escala, que não podem ser feitas em escala menor, de forma a extrair novas ideias ou criar novas formas de valor, de maneira que acabam mudando mercados, organizações, a relação entre cidadãos e os governos, dentre outros."
(TAURION, 2013)	"Big Data = volume + variedade + velocidade."
(CRAWFORD; SCHULTZ, 2014)	"Big Data é um termo generalizado e impreciso, que se refere ao uso de grandes conjuntos de dados na ciência de dados e análise preditiva."
(GEORGE et. al.2014)	"O Big data é formado por uma crescente pluralidade de fontes de informação, entre eles cliques na web, transações em dispositivos moveis, conteúdo gerado por usuários, mídias sociais, bem como conteúdo gerado intencionalmente através de redes de sensores ou transações comerciais, tais como consultas de vendas e transações de compra."
(IBM, 2014)	"Big Data é um termo utilizado para descrever grandes volumes de informações de dados e que ganha cada vez mais relevância à medida que a sociedade se depara com o aumento sem precedentes no número de informações."
FRICKÉ (2015)	"Big Data não é incompatível com a experimentação. Mas é o amigo da observação passiva: encoraja a observação passiva. Realizar pesquisas e fazer observações é uma observação passiva." "Big Data deve ser capaz de nos fornecer melhores evidências para nossas teorias."
(AMARAL, 2016)	"Big Data é o fenômeno em que dados são produzidos em vários formatos e armazenados por uma grande quantidade de dispositivos e equipamentos."
(HIJMANS, 2016)	"A evolução da era big data implica, por sua própria natureza, em falta de controle, uma vez que o volume de dados é sem precedentes, diverso em variedade e movendo-se com uma velocidade que se aproxima cada vez mais do tempo real."
(ITSRio, 2016)	"[...] o conjunto de dados cuja existência só é possível em consequência da coleta massiva de dados que se tornou possível nos últimos anos, graças à onipresença de aparelhos e sensores na vida cotidiana e do número crescente de pessoas conectadas a tais tecnologias por meio de redes digitais e também de sensores."
(PROVOST; FAWCETT, 2016)	"[...] o termo *big data* significa conjunto de dados que são grandes demais para os sistemas tradicionais de processamento e, portanto, exigem novas tecnologias para processá-los"
(DAVENPORT, 2017)	"O big data é inegavelmente grande, mas o termo é um tanto ou quanto equivocado."
(MACHADO, 2018)	"Este grade volume de dados, esta explosão de dados gerados universalmente a cada instante, é chamado de Big Data, que está expondo uma nova onda de tecnologia e arquitetura destinada a extrair valor de uma imensa variedade de dados, o que permite alta velocidade com o objetivo de capturar, descobrir e analisar estas informações importantes e valiosas no âmbito de gestão de negócios."

Fonte: elaborado pelo autor com base em: (VOLPATO; RUFINO; DIAS, 2014; GOMES, 2017).

OS PRINCÍPIOS *FAIR* COMO DIRETRIZES PARA MAXIMIZAR O USO E (RE)USO DOS DADOS DE PESQUISA:

traçando um paralelo com a gestão de documentos arquivísticos digitais

THE FAIR PRINCIPLES AS GUIDELINES TO MAXIMIZE THE USE AND (RE)USE OF THE RESEARCH OF THE DATA: tracing a parallel with the management of digital archival documents

Bárbara Carvalho Diniz[1], Guilherme Ataíde Dias[2], Alzira Karla Araújo da Silva[3]
(1) Universidade Federal da Paraíba (UFPB), Jardim Cidade Universitária, s/n – Castelo Branco, João Pessoa – PB, 58051-900, dinizbarbara6@gmail.com
(2) Universidade Federal da Paraíba (UFPB), Jardim Cidade Universitária, s/n – Castelo Branco, João Pessoa - PB, 58051-900, guilhermeataide@ccsa.ufpb.br
(3) Universidade Federal da Paraíba (UFPB), Jardim Cidade Universitária, s/n – Castelo Branco João Pessoa – PB, 58051-900, alzirakarlaufpb@gmail.com

Resumo:
Aborda o que são os princípios *FAIR*, seu surgimento e sua aplicação para maximizar o uso e reuso de dados de pesquisa. Buscou-se traçar um paralelo com a gestão de documentos arquivísticos digitais e como ambos podem contribuir na gestão de dados de pesquisa. É por isso que o objetivo principal deste artigo consiste em compreender como os princípios *FAIR* e a gestão de documentos arquivísticos digitais podem contribuir para tal gerenciamento. Metodologicamente, a pesquisa é de natureza aplicada, possui caráter exploratório, e seu procedimento técnico é de uma pesquisa bibliográfica. Foram então estudados os princípios e diretrizes do *FAIR*, como também da gestão de documentos arquivísticos digitais. Estabelece-se as semelhanças e os pontos em comum das ferramentas usadas para gerir dados e documentos digitais. A partir do paralelo traçado conclui-se que por mais que os objetos digitais (dados e documentos) possuam diferenças, existem semelhanças em relação a sua curadoria, procedimentos em comum em relação ao encontro, acesso e reuso de dados e documentos.
Palavras-chave: Princípios *FAIR*. Gestão de dados de pesquisa. Gestão de Documentos Arquivísticos Digitas.

Abstract:
It addresses what FAIR principles are, their emergence, and their application to maximize the use and reuse of research data. It sought to draw a parallel with the management of digital archival documents and how both can contribute in the management of research data. This is why the main goal of this article is to understand how FAIR principles and the management of digital records can contribute to such management. Methodologically, the research is of an applied nature, has an exploratory character, and its technical procedure is a bibliographical research. The principles and guidelines of FAIR were studied, as well as the management of digital archival documents. It establishes the similarities and points in common of the tools used to manage data and digital documents. From the parallel traced it is concluded that although the digital objects (data and documents) have differences, there are similarities with respect to their curatorship, common procedures in relation to finding, accessing and reusing data and documents
Keywords:
FAIR principles. Research data management. Digital archives document management.

1 INTRODUÇÃO

Os princípios *FAIR* foram desenvolvidos em uma conferência sobre Ciências da Vida. Nessa conferência foram propostas e aprovadas pela comunidade científica um conjunto de diretrizes que tratassem da encontrabilidade, acessibilidade, interoperabilidade e reuso da vasta quantidade de dados produzidos pela ciência. Os princípios propostos têm como objetivo fazer com que um conjunto de dados possa ser encontrável (*findable*), acessível (*acessible*), interoperável (*interoperable*) e reusável (*reusable*) (WILKINSON et. al, 2016). Os princípios *FAIR* surgiram a partir dos fenômenos da *e-science*, da ciência aberta e principalmente da necessidade do gerenciamento de dados de pesquisa. É primordial entender o que são dados de pesquisa para saber como tratá-los e preservá-los. Conhecer como surgem, suas fases e sua natureza é essencial para compreender os princípios *FAIR*. A OECD (2007) conceitua dados de pesquisa como registros, sendo eles: escores numéricos, registros textuais, imagens e sons que são utilizados como fontes primárias da pesquisa científica de forma a validar seus resultados. Um conjunto de dados de pesquisa é a constituição da representação do sujeito ou objeto investigado. Tal termo não se refere a cadernos de laboratório, análisespreliminares, rascunhos de artigos científicos, planos para pesquisas futuras, revisão por pares ou comunicações com colegas ou objetos físicos (por exemplo, amostras de laboratório, bactérias e teste em animais como ratos). Tais princípios e diretrizes estão voltados, principalmente, para os dados de pesquisa legíveis por computador, pois neste formato é que reside o potencial de melhorias na distribuição eficiente dos dados e sua aplicação para pesquisa.

Dataone (2018) ressalta que o ciclo de vida de dados compreende as fases do planejamento, coleta, asseguramento, descrição, preservação, descobrimento, integração e análise de dados. Isso tudo é cíclico, pois os dados podem ser úteis não só para a pesquisa que os gerou, mas para outras pesquisas com diferentes intenções do original. Para que a gestão de dados seja aplicada corretamente, é necessário termos um plano de gestão de dados. O planejamento então é essencial para a gestão, deve-se começar a partir do início da pesquisa, considerando como os dados serão gerenciados e como serão compartilhados futuramente. Ou seja, o plano de gestão de dados é um documento formal no qual se estabelece o compromisso de como tais dados serão tratados durante o desencadeamento da pesquisa, como também após a sua conclusão (SAYÃO; SALES, 2015). Em paralelo aos princípios *FAIR*, ciclo de vida de dados e o plano de gestão de dados, tem-se a Gestão Arquivística de Documentos Digitais. Por mais que o conceito do que são dados de pesquisa, seja diferente do significado de documento arquivístico digital, os conjuntos de dados e os documentos digitais possuem algumas semelhanças na forma em como devem ser gerenciados.

A preocupação sobre serem encontráveis, interoperáveis, acessíveis, reusáveis e preservados existem em ambos objetos digitais. Paes (2005, p.26) conceitua documento como "registro de uma informação, independentemente de sua natureza do suporte que a contém." Logo complementa que o documento de arquivo é aquele que é produzido e/ou recebido por instituição pública ou privada no exercício de suas atividades, que tem como objetivo constituir como elemento de prova ou informação. Também pode ser produzido ou recebido por indivíduos ao longo de sua existência (PAES, 2005). O Arquivo Nacional (2005, p.75) traz a seguinte definição sobre documento digital: "documento codificado em dígitos binários, acessível por meio de sistema computacional." Um dos fatores que preocupa o gerenciamento de documentos digitais é em relação a sua preservação,

visto que os meios digitais são bastante voláteis afetando questões como fragilidade do suporte, obsolescência de produtos de *software* ou programas de computador, formatos de arquivo, falta de confiabilidade e do espaço de armazenamento.

Em síntese, constata-se que apesar de diferentes, conjunto de dados e os documentos arquivísticos digitais possuem características em comum. Ambos os gestores procuram exercer cuidados em relação a sua curadoria. Este trabalho procura explorar como é feito o uso e reuso de dados de pesquisa a partir do princípio *FAIR* e como é feita a gestão de documentos arquivísticos digitais, mostrando então suas semelhanças e como estes podem contribuir para a gestão dos dados de pesquisa.

2 OBJETIVOS

Objetiva-se compreender como os princípios *FAIR* e a gestão de documentos arquivísticos digitais podem contribuir para o gerenciamento de dados científicos. Em relação aos seus objetivos específicos tem como intenção: conhecer os princípios orientadores *FAIR*; entender as diretrizes e princípios sobre gestão de documentos arquivísticos digitais e; estabelecer as semelhanças entre os princípios *FAIR* e a gestão de documentos arquivísticos digitais.

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa classificada como de natureza aplicada, pretende gerar conhecimentos sobre os princípios *FAIR* associados a gestão de documentos arquivísticos digitais, visando sua aplicação na prática para solucionar os problemas de gerenciamento de dados. Caracteriza-se como exploratória, pois traz à tona o problema do gerenciamento de dados de pesquisa, e levanta estudos bibliográficos para a construção de soluções que possam colaborar com a gestão e preservação dos dados de pesquisa. Em relação aos seus procedimentos técnicos é classificada como uma pesquisa bibliográfica, porque "é desenvolvida com base em material já elaborado, constituído principalmente de livros e artigos científicos." (GIL, 2002, p.44).

4 DISCUSSÕES

Primeiramente abordaremos sobre o que são os princípios *FAIR* e como se desenvolve a gestão de documentos arquivísticos digitais. Trata da comparação entre o gerenciamento de dados de pesquisa e o de documentos arquivísticos digitais.

4.1 Princípios *FAIR*

No início deste trabalho foi mencionado que os princípios *FAIR* surgiram e foram desenvolvidos em uma conferência de Ciências da Vida. Os pesquisadores presentes propuseram

e elaboraram quatro princípios para maximizar o uso e reuso dos dados, são eles: *Findability* (encontravéis), *Acessible* (acessível), *Interoperable* (interoperável) e *Reusable* (reusável). Wilkinson et al (2016) relata que tais princípios servem para orientação dos produtores e editores a como navegar em torno dos obstáculos dos dados. Estes princípios ajudam a maximizar o valor adquirido nas publicações acadêmicas contemporâneas em formato digital. Em resumo, todos os objetos de pesquisa digitais dos dados para a análise de conduta acadêmica, também devem aplicar os princípios *FAIR*, pois todos os componentes do processo da pesquisa deverão estar disponíveis para garantir a transparência, reprodução e reuso dos dados. A iniciativa *FORCE* 11 (2018) publicou diretrizes para seguir os princípios *FAIR*. A seguir um quadro especificando o que é necessário para se cumprir estes princípios.

Quadro 1 – Diretrizes dos princípios *FAIR*

PRINCÍPIO	**ESPECIFICAÇÕES**
Encontrável (*Findable*)	Reencontrável em qualquer ponto no tempo; dados digitais devem ser persistentes, com ênfase em seus metadados; minimamente conter metadados acionáveis por máquina que permitam distingui-lo de outros objetos de dados; uso de identificador único e persistente.
Acessível (acessible)	Devem ser acessíveis, pois devem ser obtidos tanto por máquinas, quanto por humanos; mediante autorização apropriada; por meio de protocolo bem definido; assim máquinas e humanos serão capazes de julgar se o objeto é acessível.
Interoperáveis *(interoperable)*	Metadados devem ser acionados por máquinas; os (meta) formatos de dados devem utilizar vocabulários compartilhados e ontologias; os metadados dentro dos objetos de dados devem ser sintaticamente reproduzíveis e semanticamente acessíveis.
Reusáveis (*reusable*)	Os objetos de dados devem estar em conformidade com os princípios da encontrabilidade (encontrável) e interoperável; metadados devem ser ricos para que com o mínimo de esforço humano possam ser ligados ou integrados com outras fontes de dados; deve-se referir as suas fontes com metadados ricos para permitir a citação apropriada.

Fonte: Adaptado do *FORCE* 11 (2018)

As diretrizes são claras quando enfatizam o uso de metadados e identificadores persistentes, autorizações e licenças para publicação e acessibilidade, como também o uso de vocabulários compartilhados e ontologias para o acesso em geral aos dados.

4.2 Gestão de Documentos Arquivísticos Digitais

O Conselho Nacional de Arquivos (2011) relata que a gestão de documentos arquivísticos trata de procedimentos e técnicas da produção, tramitação, uso, avaliação e arquivamento dos

documentos produzidos e recebidos por entidades ou indivíduos no decorrer de sua existência em fase corrente e intermediária. Os documentos então podem passar por até três fases, são elas: corrente, intermediária e permanente. A fase corrente se refere aos documentos usados com frequência e que são mantidos perto do setor que os produziu. Na fase intermediária os documentos são pouco utilizados e são mantidos longe de seus produtores aguardando prazos prescricionais, em que dependendo de avaliação pode ser eliminado ou recolhido para o arquivo permanente. Por fim, a fase permanente custodia os documentos que devem ser preservados por seu valor histórico, científico, cultural e informativo. Com o advento das Tecnologias da Informação, os documentos começaram a ser produzidos e armazenados em meio eletrônico. Para isso, é necessário que a gestão de documentos possa também ser aplicada em meio digital. A seguir o Quadro 2 demonstra os procedimentos e ações da gestão de documentos arquivísticos digitais elaboradas pelo Conselho Nacional de Arquivos (2011).

Quadro 2 – Gestão de documentos arquivísticos digitais (procedimentos e ações)

PROCEDIMENTO	AÇÕES
Captura Consiste no procedimento que declara um documento como documento arquivístico, este procedimento incorpora o documento ao sistema.	Registro, classificação, indexação, atribuição de restrição de acesso e arquivamento.
Avaliação Uma atividade essencial para a gestão de documentos, pois por meio da avaliação é possível racionalizar o grande número de documentos nas fases corrente e intermediária, fazendo então que só os documentos essenciais cheguem a fase permanente.	Retenção dos documentos, eliminação, transferência e recolhimento.
Pesquisa, localização e apresentação dos documentos O sistema deve possuir funções para recuperação e acesso aos documentos e as informações contidas neles, como forma de facilitar o dia a dia das atividades das instituições.	A partir de parâmetros definidos que são selecionadas entre as informações coletadas no momento do registro do documento e entre os metadados associados a ele.
Segurança: controle de acesso, trilhas de auditoria e cópias de segurança O sistema deve conter controles de acesso e procedimentos de segurança que possam garantir a integridade dos documentos.	Controle de acesso, uso e rastreamento, trilha de auditoria, cópias de segurança e segurança da infraestrutura.
Armazenamento O armazenamento deve garantir autenticidade e o acesso aos documentos pelo tempo estabelecido pela tabela de temporalidade (define o tempo de guarda dos documentos, seja por tempo determinado ou guarda permanente) e sua destinação.	Fatores mais importantes para a seleção das opções de armazenamento: volume e estimativa dos documentos, segurança dos documentos, características do suporte e do ambiente, frequência de uso e custo relativo das opções de armazenamento dos documentos.
Preservação Devem se manter acessíveis e utilizáveis pelo tempo necessário para sua guarda. Deve ter garantido sua longevidade, funcionalidade e seu acesso contínuo.	Prevenção de obsolescência tecnológica e de danos físicos ao suporte que se encontra armazenado, técnicas como migração, *refreshing*, conversão, emulação, encapsulamento e preservação a tecnologia.

Fonte: Conselho Nacional de Arquivos (2011).

4.3 Pontos em comum

No apêndice A é demonstrado detalhadamente em um quadro quais os pontos em comum entre os princípios *FAIR* e a gestão de documentos arquivísticos digitais. Em resumo, notar que a gestão de dados de pesquisa a partir do uso dos princípios *FAIR* e a gestão de documentos digitais arquivísticos apresentam algumas diferenças em sua forma de administrar seus objetos digitais, contudo, ambas partilham de ferramentas e ações semelhantes na forma de recuperar, acessar e reusar suas informações.

5 CONSIDERAÇÕES FINAIS

Os princípios *FAIR* trazem como pilares serem encontráveis, acessíveis, interoperáveis e o reutilizáveis como preocupações da gestão de dados de pesquisa. Estes princípios focam potencialmente em administração para o uso e reuso dos dados, não abordando a questão da preservação, que também é uma preocupação da gestão dos dados de pesquisa. A gestão de documentos arquivísticos digitais tem como prerrogativas não apenas a administração, mas assuntos que envolvem a política institucional e preservação dos documentos. Não é intenção fazer uma crítica aos princípios *FAIR*, pois estes foram criados para maximizar o uso e reuso de dados. A preservação aparentemente não é seu foco. Contudo, os princípios *FAIR* aliados aos princípios da gestão de documentos arquivísticos digitais demonstram qualidades essenciais na administração e preservação de dados de pesquisa. Quem tende a ganhar com isto é a ciência, pois terá acesso e preservação para o desenvolvimento presente e futuro.

REFERÊNCIAS

ARQUIVO NACIONAL (BRASIL). **Dicionário brasileiro de terminologia arquivística**. Rio de Janeiro: Arquivo Nacional, 2005. 232 p.

CONSELHO NACIONAL DE ARQUIVOS (BRASIL). **e-ARQ Brasil**: Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos/ Câmara Técnica de Documentos Eletrônicos. 1.1 versão. Rio de Janeiro: Arquivo Nacional, 2011.136p. Disponível em: <http://www.siga.arquivonacional.gov.br/imag es/publicacoes/e-arq.pdf>. Acesso em: 24 set. 2018.

COMISSÃO EUROPEIA. **Guidelines on FAIR Data Management in Horizon 2020**. Versão 3.0. [S.l.: s.n], 2016. 12 p. Disponível em: <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020- hi-oa-data-mgt_en.pdf>. Acesso em: 24 set. 2018.

DATAONE. **Data Life Cycle**. 2018. Disponível em: <https://www.dataone.org/data-life-cycle>. Acesso em: 24 set. 2018.

FORCE 11**. Guiding principles for Findable, Acessible, Interoperable and Re- usable data publishing version b.1.0**. 2018. Disponível em: <https://www.force11.org/fairprinciples>. Acesso em: 24. set. 2018.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002. 176 p.

OECD. **OECD Principles and Guideliness for Acess to Research Data from Public Funding**. Paris: OECD Publications, 2007, 24p. Disponível em: <http://www.oecd.org/science/sci-tech/38500813.pdf>. Acesso em: 24 set. 2018.

PAES, Marilena Leite. **Arquivo**: teoria e prática. 3. ed. Rio de Janeiro: FGV Editora, 2005. 228p.

SAYÃO, Luiz Fernando; SALES, Luana Farias. **Guia de Gestão de Dados de Pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN/IEN, 2015. 90p. Disponível em: <http://carpedien.ien.gov.br/bitstream/ien/162 4/1/GUIA_DE_DADOS_DE_PESQUISA.pdf>. Acesso em: 24 set. 2018.

WILKINSON, Mark D. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. **Scientific Data**. Nature. 2016. Disponível em: <https://www.nature.com/articles/sdata20161 8#article-info>. Acesso em: 24 set. 2018.

Apêndice A – SEMELHANÇAS ENTRE OS PRINCÍPIOS ***FAIR*** E A GESTÃO DE DOCUMENTOS ARQUIVÍSTICOS DIGITAIS

Princípios *FAIR*	**Gestão de documentos arquivísticos digitais**	**Pontos em comum**
Dados **encontráveis**	**Encontrabilidade** de documentos.	**Pontos em comum sobre encontrabilidade.** (serem encontráveis)
Encontrável a partir dos metadados, uso de identificador persistente, uso de palavras- chave, nomenclatura apropriada.	Uso de identificador, descrição informativa = metadados, recursos de indexação, uso de palavras- chave, documentos encontrados a partir dos metadados.	Uso de identificador, encontrável a partir de metadados, uso de palavras-chave.
Dados **acessíveis**	**Acessibilidade** de documentos.	**Ponto em comum sobre acessibilidade.**
Mediante autorização apropriada, protocolo bem definido, controle de acesso, especificar onde se depositam os dados, metadados, documentação e códigos.	Controle de acesso dos documentos, atributos de segurança, administrador do sistema determina quem pode ter acesso aos documentos.	Acesso mediante a autorização apropriada, controle de acesso.
Dados **interoperáveis**	**Interoperabilidade** de documentos.	**Pontos em comum sobre interoperabilidade.**
Metadados acionados por máquinas, metaformatos utilizam vocabulários controlados e ontologias, metadados dentro dos objetos de dados deve ser reproduzíveis e acessíveis, especificar dados, vocabulários, padrões e metodologias deverá seguir.	Consulta, recuperação, importação e exportação de documentos e seus metadados, operações de interoperabilidade devem respeitar a legislação vigente e a política de segurança, capaz de interoperar com outros sistemas por meio de padrões abertos.	Não foram encontrados pontos em comum neste aspecto.
Dados **reusáveis**	**Reuso** de documentos.	**Pontos em comum sobre reuso.**
Devem estar em conformidade com os princípios de encontrabilidade (encontráveis) e interoperabilidade (interoperáveis), bem descritos para que possam ser integrados a outros dados, referência a fonte e metadados, processo de garantia de qualidade de dados, cumprimento de tempo da validade de dados.	Os documentos arquivísticos devem possuir valores de autenticidade, confiabilidade, organicidade, acessibilidade e unicidade. No plano de classificação e destinação deve estar atribuído o tempo de guarda do mesmo para fins administrativos ou históricos.	Deve ser estabelecido o tempo de guarda e validade dos dados/documentos. Garantia de qualidade, confiabilidade da informação.

Fonte: Adaptado do Conselho Nacional de Arquivos (2011), Comissão Europeia (2016) e *Force* 11 (2018)

PADRÕES DE METADADOS PARA DESCRIÇÃO DE DADOS:
panorama dos repositórios de dados na América Latina

METADATA STANDARDS FOR DATA DESCRIPTION:
overview of Data Repositories in Latin America

Felipe Augusto Arakaki[1], Ana Carolina Simionato[2], Paula Regina Ventura Amorim Gonçalez[3], Plácida Leopoldina Ventura Amorim da Costa Santos[4]
(1) Instituto Federal de São Paulo (IFSP), Câmpus Itapetininga, Av. João Olímpio de Oliveira, 1561 - Vila Asem, Itapetininga - SP, 18202-000, felipe.arakaki@ifsp.edu.br.
(2) Universidade Federal de São Carlos (UFSCar), Rodovia Washington Luis, km 235 - São Carlos - SP, 13565-905, Departamento de Ciência da Informação, acsimionato@ufscar.br;
(3) Universidade Federal do Espírito Santo (UFES), Av. Fernando Ferrari, 514 - Goiabeiras, Vitória - ES, 29075-910, Departamento de Biblioteconomia, paula.goncalez@ufes.br,
(4) Universidade Estadual Paulista (UNESP), Av. Hygino Muzzi Filho, 737, Mirante, Marília, SP, 17.525-900, Programa de Pós-Graduação em Ciência da Informação, placida@marilia.unesp.br.

Resumo:
Os repositórios de dados de pesquisa são ambientes que oportunizam aos dados, produtos de pesquisa, serem acessados, compartilhados, utilizados e reusados, visto que, tais ambientes possibilitam a organização, o armazenamento e o acesso aos dados em diferentes formatos. A preocupação com a representação e recuperação dos conjuntos de dados é recorrente entre profissionais que tem a informação como objeto de trabalho e pesquisa. Assim, questiona-se quais são os padrões de metadados utilizados na representação dos dados científicos disponibilizados nos repositórios de dados na América Latina? Nesse contexto, o trabalho possui como objetivo analisar a formalização dos padrões de metadados utilizados para a descrição do conjunto de dados de pesquisa no âmbito dos repositórios de dados da América Latina registrados no *Registry of Research Data Repositories* (*re3data*). A pesquisa é de natureza teórico aplicada, com uma abordagem qualitativa no que se refere a representação da informação nos repositórios digitais de dados, para tanto optou-se pela pesquisa exploratória. Como resultado foi feito o mapeamento dos metadados dos 29 repositórios de dados científicos identificados no *re3data*. Ressalta-se que o reconhecimento para dados científicos e seu compartilhamento tem potencializado a criação de novas plataformas, *softwares* de gerenciamento e diferentes possibilidades para a descrição dos conjuntos de dados.
Palavras-chave: Descrição de conjuntos de dados. Metadados. Repositório de dados.

Abstract:
Research data repositories are environments that provide data, research products, are accessed, shared, used and reused, since such environments enable the organization, storage and access to data in different formats. The concern with the representation and retrieval of datasets is recurrent among professionals who have information as object of work and research. Thus, it is questioned what metadata patterns are used in the representation of the scientific data available in the data repositories in Latin America? In this context, the work aims to analyze the formalization of the metadata standards used to describe the set of research data within the Latin American data repositories registered in the Registry of Research Data Repositories (re3data). The research is of an applied theoretical nature, with a qualitative approach regarding the representation of the information in the digital repositories of data, for which we have opted for the exploratory research. As a result, the mapping of the metadata of the 29 research data repositories identified in the survey was done. It should be emphasized that the recognition of scientific data and their sharing has made possible the creation of new platforms, management software and different possibilities for the description of datasets.
Keywords: Dataset description. Metadata. Data repository.

1 INTRODUÇÃO

A preocupação para o efetivo gerenciamento de dados oriundos de pesquisas está sendo amplamente discutido pelas principais agências de fomento à pesquisa, visando o progresso científico além da contemporização de uma Ciência Aberta à todos os campos do conhecimento.

Esse movimento, torna-se salutar na concepção dos Repositórios de Dados. Os Repositórios de Dados de Pesquisa, se configuram como ambientes no auxílio a pesquisadores no que diz respeito à gerência, à disponibilização e ao acesso aos dados científicos, ações fundamentais, quando este ambiente permite e assegura o compartilhamento, o acesso e a reutilização dos dados, ações que reduzem sobremaneira o tempo e os gastos com nova coleta, como lhes chancelar validade diante a reprodução e replicação da pesquisa.

Vale destacar que os dados de pesquisa são definidos como:

> "[...] dados digitais são uma parte (descritiva) ou o resultado de um processo de pesquisa. Este processo abrange todas as etapas da pesquisa, que vão desde a geração de dados de pesquisa, que podem ser em um experimento nas ciências, um estudo empírico nas ciências sociais ou observações de fenômenos culturais, até a publicação dos resultados da pesquisa." (PAMPEL et al., 2013, p. 1, tradução nossa).

A disponibilização dos dados científicos para o acesso, usos e reuso, objetivo primeiro dos repositórios de dados, requer um planejamento e gerenciamento eficiente, segundo Simionato (2017) essas ações devem ser iniciadas desde a confecção do *Data Management Plan* (DMP) pelo pesquisador no depósito dos dados durante toda sua execução até sua finalização.

Para que os dados possam ser disponibilizados à comunidade científica se faz necessário apresentar requisitos padronizados e validados segundo normas internacionais e assegurar uma infraestrutura adequada para preservação digital.

Nesse sentido, parte desta infraestrutura equivale a composição dos metadados por meio de seus padrões. Assim, a representação e o armazenamento dos dados é uma preocupação recorrente do profissional que tem a informação como seu objeto de trabalho.

Há grande relação entre dados e metadados. Suas definições podem ser tratadas como muito similares. Jeffery et al. (2014, não paginado, tradução nossa) explicam que a percepção pode variar a partir do contexto da análise. Por exemplo, para um pesquisa que faz uma busca em uma base de dados, ele utiliza metadados para descobrir um livro ou artigo, já para quem gerencia o sistema, a base de dados pode ser utilizada como dados para analisar as coleções por assunto, por editora, por ano etc.

Entre os principais padrões de metadados para descrição de objetos digitais está o *Dublin Core* (DC). O DC surgiu em 1995, com uma proposta de a partir de um núcleo reduzido, representar qualquer documento na *Web*. (ARAKAKI, ALVES, SANTOS, 2018).

Nesse contexto, o armazenamento, representação são fundamentais para recuperação e acesso à informação. Rodrigues et al. (2010) aponta que os repositórios de dados provêem serviços dirigidos à quem deposita e aos provedores do sistema que são como uma extensão dos repositórios digitais, porém afirma que: "Nos repositórios de dados pode-se ir além desta visão de repositórios

de objetos, uma vez que cada conjunto de dados possui características próprias e por isso pode requerer um tratamento diferenciado" (RODRIGUES et al., 2010, p. 23).

Foram criados diversos tipos de repositórios de dados, como os repositórios temáticos que buscam reunir dados de uma área do conhecimento. Há uma tendência ainda, para os repositórios de dados institucionais, quando as Instituições de Ensino disponibilizam os dados oriundos de pesquisa para comunidade e ainda há possibilidade de repositórios relacionados à grupos de pesquisa conforme explorado por Vidotti et al. (2017).

Sales e Sayão (2015, p. 28) ao discorrerem sobre a implantação de repositórios de dados e curadoria digital de dados pontuam que "[...] as exigências sobre o nível de descrição e de atribuição de metadados devem ser identificadas desde o começo de seu projeto e revistas ao longo de vida dos seus dados". Os autores ainda complementam afirmando que essa é a essência da curadoria dos dados, ou seja, os metadados deverão assegurar a disponibilização dos dados para seu uso e reuso.

Nesse cenário, o crescimento do número de repositórios de dados nos leva a seguinte indagação: quais são os padrões de metadados utilizados na representação dos dados científicos disponibilizados nos repositórios de dados na América Latina?

Santos e Alves (2013) ao discorrerem sobre os padrões de metadados apontam que suas estruturas padronizadas representam um conteúdo informacional intencionam sua recuperação e acesso, ou seja, os padrões de metadados são "[...] um conjunto estruturado, padronizado, codificado e pré-determinado de elementos de metadados que serão utilizados na representação descritiva dos recursos informacionais, aplicações e ou compartilhamento de dados entre sistemas (ALVES; SANTOS, 2013, p. 13).

Nesse cenário, esse trabalho tem como objetivo analisar a formalização dos padrões de metadados utilizados para a descrição do conjunto de dados de pesquisa no âmbito dos repositórios de dados da América Latina registrados no R*egistry of Research Data Repositories* (re3data).

2 PROCEDIMENTOS METODOLÓGICOS

O trabalho apresenta os resultados de uma abordagem qualitativa, referente à representação da informação nos repositórios de dados científicos. Para isso, optou-se pela realização de uma pesquisa exploratória e documental da literatura, nacional e internacional, sobre a temática e da descrição de conjunto de dados em repositórios de dados.

Durante a fase de identificação e verificação dos resultados, optou-se pela cobertura geográfica da América Latina. A base escolhida para a busca dos repositórios foi a *re3data.org* (disponível em: https://www.re3data.org/) sendo os resultados recuperados até o dia 04 de agosto de 2018.

A *re3data.org* configura-se como um registro global de repositórios de dados de pesquisa que abrange repositórios de dados de pesquisa de diferentes disciplinas acadêmicas. O filtro utilizado foi a cobertura geográfica, no caso a América Latina, e os resultados apresentaram 29 repositórios de dados, que serão abordados no próximo tópico.

3 RESULTADOS

Dos 29 repositórios identificados, foram localizados onze (11) repositórios no México, oito (8) repositórios no Brasil, Argentina, Colômbia e Panamá foram localizados dois (2) repositórios cada e Chile, El Salvador e Peru foi localizado um (1) repositório cada. Importante relatar que até o dia 4 de agosto de 2018, nos países: Bolívia, Costa Rica, Cuba, Equador, Guatemala, Haiti, Honduras, Nicarágua, Paraguai, República Dominicana, Uruguai e Venezuela não foram encontrados nenhum repositório cadastrado na *re3data.org*.

Durante o levantamento, foi observado que cinco (5) repositórios utilizam o software *Dspace*, três (3) repositórios utilizam *software DataVerse*, dois (2) repositórios utilizam *Drupal,* um (1) repositório utiliza o *Metacat*, e um repositório informou que usa a denominação de "outro", mas não especificou o *software* utilizado. Entre os repositórios analisados, um considerou o *MySQL* como sistema utilizado e 16 repositórios apontaram que o sistema é desconhecido no contexto do *re3data* ou não informaram. O uso de um *software* adequado para organização e tratamento dos dados influencia diretamente nas possibilidades de recuperação, armazenamento, representação da informação e o estabelecimento de padrões de metadados.

Nesse contexto os repositórios da América Latina, os repositórios: *Repositorio Institucional USIL* (Peru), *Repositorio Institucional UCASAL* (Argentina), Base de Dados Científicos da Universidade Federal do Paraná (Brasil), CEDAP *Research Data Repository - research data* (Brasil) adotam o *Dublin Core*. Destaca-se que sua composição básica possui 15 elementos que são opcionais e repetíveis, além da possibilidade de expandir a quantidade de elementos e refinamentos.

Em contrapartida, há diversos outros padrões específicos para descrição de dados no contexto dos repositórios de dados. Conforme o levantamento realizado, três (3) repositórios (CIAT *Dataverse* - Colômbia; IBICT *Dataverse Network* - Brasil; e CIMMYT *Research Data & Software Repository Network* - Mexico) utilizam um padrão internacional para descrição de dados científicos, o *Data Documentation Initiative* (DDI). Os metadados do DDI permitem representar o conjunto de dados e gerenciar diferentes estágios no ciclo de vida dos dados de pesquisa, como conceituação, coleta, processamento, distribuição, descoberta e arquivamento. (DATA DOCUMENTATION INITIATIVE, 2018).

O padrão *Darwin Core* é utilizado no Portal de datos de Biodiversidad (Argentina). O *Darwin Core* está estruturado com base no *Dublin Core*, XML e RDF Schema e inclui um glossário de termos tendo como objetivo facilitar o compartilhamento de informações sobre diversidade biológica. (DARWIN CORE TASK GROUP, 2014).

No contexto dos dados ecológicos, o *PPBio Data Repository* (Brasil) adotou o *Ecological Metadata Language* (EML), que é uma especificação de metadados que está na versão 2.1.1, implementado como uma série de tipos de documentos XML que podem ser usados de maneira modular e extensível. (KNOWLEDGE NETWORK FOR BIOCOMPLEXITY, 2015).

Para descrição de dados geoespaciais, o repositório *Integrated Taxonomic Information System* (ITIS) do México optou por utilizar o *Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata* (FGDC/CSDGM) que estabelece e implementa orientações para descrição do conteúdo, da qualidade e da transferência de dados geoespaciais. Entre os tipos de materiais

estão mapas, arquivos de Sistemas de Informações Geográficas (GIS), imagens e outros recursos. (FEDERAL GEOGRAPHIC DATA COMMITTEE, 2014).

Para o intercâmbio de dados no âmbito da área da astronomia o *Flexible Image Transport System* (FITS) foi adotado pelo repositório *Gran Telescopio CANARIAS Public Archive*, do México. O FITS é um padrão recomendado pela NASA e pela *International Astronomical Union* e usado para o transporte, análise e armazenamento de arquivos de conjuntos de dados científicos, pode apresentar matrizes multidimensionais como espectros 1D, imagens 2D, 3D e mais os cubos de dados. (NASA, 2017).

Entre os repositórios analisados, a ISO 19115 que define o esquema para descrever informações e serviços geográficos por meio de metadados é utilizado no *International Ocean Discovery Program* (Brasil). O padrão da ISO fornece informações sobre a identificação, a extensão, a qualidade, os aspectos espaciais e temporais, o conteúdo, a referência espacial, a representação, a distribuição e outras propriedades dos dados e serviços geográficos digitais. (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2014).

O *Access to Biological Collection Data* (ABCD) é utilizado pelo repositório internacional GLOBE, com abrangência no Brasil. O ABCD é um padrão que abrange a descrição de dados de espécimes de coleções de história natural e de espécies em geral. (BOTANIC GARDEN AND BOTANICAL MUSEUM BERLIN-DAHLEM, 2016).

Diante o levantamento, foi observado que mais da metade dos repositórios analisados, dezesseis (16), não informaram se utilizam algum padrão de metadados ou se optaram pelos metadados já identificados como descritores do *software* aderido.

4 CONSIDERAÇÕES FINAIS

Durante o levantamento realizado, constatou-se que grande parte dos repositórios da América Latina não utilizam ou não informaram os padrões de metadados adotados, isto é, como já apresentado nos resultados 16 repositórios não referenciam à padrões.

O padrão de metadados mais adotado entre os 29 repositórios de dados é o *Dublin Core*, em razão da utilização do *software Dspace* que já faz parte da instalação da plataforma.

Em compensação, alguns repositórios utilizam metadados específicos da área do repositório, possibilitando uma descrição mais exaustiva das informações além de permitir a interoperabilidade entre sistemas.

Outro ponto de destaque, é que os princípios de descrição dos conjuntos de dados são fundamentados nos mesmos princípios de descrição de outros tipos de recursos informacionais, sendo que propriedades de clareza, precisão, lógica e integridade devem ser equalizadas para todos os tipos de recursos, diferenciando-se apenas, em especificidades relativas à forma e formato.

O reconhecimento para a importância dos dados científicos, bem como o seu compartilhamento em repositórios vêm se potencializando com a criação de novas plataformas, *softwares* para o gerenciamento e outras possibilidades para a descrição de conjuntos de dados, entretanto, ainda carece de pesquisas e desenvolvimento.

REFERÊNCIAS

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

ARAKAKI, F. A.; ALVES, R.C.V.; SANTOS, P.L.V.A. da C. Dublin Core: state of art (1995 to 2015). **Informação & Sociedade**: Est., João Pessoa, v. 28, n. 2, p. 7-20, maio/ago. 2018. Disponível em: <http://www.periodicos.ufpb.br/ojs2/index.php/ies/article/view/38012/pdf>. Acesso em: 19 nov. 2018.

BOTANIC GARDEN AND BOTANICAL MUSEUM BERLIN-DAHLEM. **Access to Biological Collection Data schema**. 2016. Disponível em: <http://www.biocase.org/products/schema_repository/>. Acesso em: 6 ago. 2018.

DARWIN CORE TASK GROUP. **Darwin Core**. 2014. Disponível em: <http://rs.tdwg.org/dwc/index.htm>. Acesso em: 6 ago. 2018.

DATA DOCUMENTATION INITIATIVE. **Welcome to the Data Documentation Initiative**. 2018. Disponível em: <http://www.ddialliance.org/>. Acesso em: 6 ago. 2018.

FEDERAL GEOGRAPHIC DATA COMMITTEE. **Geospatial Metadata**. 2014. Disponível em: <https://www.fgdc.gov/metadata>. Acesso em: 6 ago. 2018.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 19115-1:2014** - Geographic information -- Metadata -- Part 1: Fundamentals. Disponível em: <https://www.iso.org/standard/53798.html>. Acesso em: 6 ago. 2018.

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: 17 out. 2015.

KNOWLEDGE NETWORK FOR BIOCOMPLEXITY. **Ecological Metadata Language (EML)**. 2015. Disponível em: <https://knb.ecoinformatics.org/#external//emlparser/docs/index.html>. Acesso em: 6 ago. 2018.

NASA. **FITS Support Office**. 2017. Disponível em: <https://fits.gsfc.nasa.gov/>. Acesso em: 6 ago. 2018.

PAMPEL, H. et al. Making research data repositories visible: The re3data. org registry. **PloS one**, v. 8, n. 11, p. e78080, 2013. Disponível em: <http://bit.ly/2vHTqGA>. Acesso em: 07 out. 2018.

RODRIGUES, E. et al. **Os repositórios de dados científicos:** estado da arte. 2010. Disponível em: <http://projeto.rcaap.pt/index.php?option=com_remository&Itemid=2&func=startdown&id=271&lang=pt>. Acesso em: 15 set. 2018.

SAYÃO L. F.; SALES, L. F. Guia de gestão de dados de pesquisa para bibliotecários e pesquisadore. Rio de Janeiro : CNEN/IEN, 2015. Disponível em:< http://carpedien.ien.gov.br/bitstream/ien/1624/1/GUIA_DE_DADOS_DE_PESQUISA.pdf> Acesso em: 08 mar. 2017.

SIMIONATO, A. C. Mapeamento dos metadados para dados científicos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18. 2017, Marília. **Anais…**: Marília: UNESP, 2015. Disponível em <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/view/563> : Acesso em: 02 de out. 2018.

VIDOTTI, S. A. B. G. et al. Repositório de dados de pesquisa para grupo de pesquisa: um projeto piloto. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18, 2017, Marília. **Anais...** Marilia: UNESP, 2017.Disponível em: <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/388/932>. Acesso em: 17 set. 2018

POLÍTICA DE COMPARTILHAMENTO DE DADOS CIENTÍFICOS:
a adoção nos periódicos da Ciência da Informação

SCIENTIFIC DATA SHARING POLICY:
adoption in Information Science journals

Renata Lemos dos Anjos[1], Marynice de Medeiros Matos Autran[(2)]
(1) Universidade Federal da Paraíba, João Pessoa - PB, renatalemosdosanjos@gmail.com.
(2) Universidade Federal da Paraíba, João Pessoa - PB, marynice.autran@gmail.com.

Resumo:
Desde 1665, a comunicação científica continua prioritariamente direcionada à publicação dos resultados em periódicos científicos. Esses resultados provêm de análises realizadas em nas investigações científicas por meio dos dados que são coletados para a pesquisa, ou gerados com a pesquisa. Com o desenvolvimento das Tecnologias Digitais de Informação e Comunicação, a comunicação científica passou por várias transformações. Atualmente, depara-se com um novo paradigma: a do compartilhamento de dados. A presente pesquisa exploratória e documental, investigou se os periódicos da área de Ciência da Informação possuem uma política de compartilhamento de dados científicos. Foram selecionados periódicos com o CiteScore maior ou igual a 0.50. Dos quinze periódicos analisados, 53% apresentam políticas e diretrizes para o compartilhamento de dados, seja de forma obrigatória ou flexível. Concluiu-se que a prática do compartilhamento de dados por parte dos periódicos em Ciência da Informação, encontra-se nos seus em fase inicial e apresenta perspectivas de breve consolidação.
Palavras-chave: Política de compartilhamento de dados. Periódicos. Ciência da informação.

Abstract:
Since 1665, scholarly communication has continued to focus primarily on publication of results in scientific journals. These results come from analyses carried out in scientific investigations through data collected for research, or generated through research. With the development of Digital Information and Communication Technologies, scholarly communication has undergone several transformations. Nowadays, we are seeing a new paradigm: the data sharing. This exploratory and documentary research investigated whether the Information Science journals have a policy of sharing scientific data. Journals were selected with the CiteScore greater than or equal to 0.50. Of the fifteen journals analyzed, 53% present policies and guidelines for data sharing, whether in a mandatory or flexible way. The practice of data sharing by Information Science journals is in initial phase and presents prospects of brief consolidation.
Keywords: Data sharing policy. Journals. Information science.

1 INTRODUÇÃO

Em 1665 foram publicados os primeiros periódicos científicos, na Inglaterra, o *Philosophical Transactions of the Royal Society* e na França, o *Journal des Sçavans*. Desde então, a Comunicação Científica - CC, direciona à publicação, prioritariamente, os resultados de pesquisas para publicação em periódicos científicos (ANJOS, 2016). Esses resultados provêm de análises realizadas no curso das investigações, por meio dos dados coletados para a pesquisa, ou gerados a partir desta.

Com a emergência das Tecnologias Digitais de Informação e Comunicação – TDIC, tanto a própria ciência, quanto a CC passaram por transformações, afirmando o seu caráter evolutivo, mutável e dinâmico no que concerne ao seu modo de divulgar e comunicar a ciência (TARGINO, 2000). Essas transformações referem-se à adaptação ao cenário eletrônico, de forma que os processos de publicação dos periódicos científicos foram alterados, tanto nos moldes de avaliação e comunicação entre autores, editores, pareceristas e leitores, quanto nos moldes de disseminação e acesso às publicações (NASSI-CALÒ, 2014).

Diante de tal realidade, a CC encontra-se imersa no fenômeno da *e-Science*, considerado como o quarto paradigma científico (GRAY, 2007). Em linhas gerais *a e-Science* consiste no uso e no intensivo compartilhamento de dados provenientes de investigações científicas, fazendo com que se tornem matéria prima para outras investigações, por meio da sua reutilização (SAYÃO; SALES, 2014).

Nesta perspectiva, fundamentando-se no compartilhamento intensivo de dados e em seu posterior reuso, periódicos das mais variadas áreas do conhecimento estão dando especial atenção aos dados resultantes de investigações científicas, bem como solicitando aos autores que disponibilizem, juntamente com o artigo final, o conjunto de dados utilizados na pesquisa, de forma a incentivar políticas de compartilhamento de dados científicos.

Em concordância com Nassi-Calo (2014) ao referir que "[...] quanto mais pesquisadores disponibilizarem abertamente seus dados em repositórios de acesso aberto, maior a probabilidade de que outros possam replicar seu trabalho, com benefícios evidentes para todos". Nesta concepção, a política de compartilhamento de dados científicos, atualmente, vem sendo incentivada. Percebemos que um mesmo conjunto de dados possibilita várias vertentes de exploração, e podem também validar ou revogar uma determinada pesquisa e artigo científico. O incentivo a essas práticas de compartilhamento de dados, faz a comunicação científica desenvolver-se cada vez mais, além de aumentar a qualidade da ciência.

2 OBJETIVOS

Esta pesquisa busca analisar a adoção da política de compartilhamento de dados nos periódicos da área da Ciência da Informação com CiteScore maior ou igual a 0.50 e verificar como ocorre esse compartilhamento/disponibilização, se de maneira obrigatória ou flexível.

3 PROCEDIMENTOS METODOLÓGICOS

Trata-se de uma pesquisa exploratória e documental, que buscou identificar nas *homepages* dos periódicos informações sobre a adoção de políticas de compartilhamento de dados.

Foram selecionados os periódicos da área de Library and Information Sciences (Biblioteconomia e Ciência da Informação) de acesso aberto, que estão indexados na base de dados bibliográfica Scopus. Essa base apresenta, entre outras métricas, o CiteScore – que mede a influência de determinado periódico em sua área por meio do quociente entre a quantidade de vezes que o periódico foi citado em determinado ano, no caso 2017, e a quantidade de documentos (artigos) disponíveis nos últimos três anos (2014, 2015, 2016).

Como critério de seleção, escolhemos aqueles periódicos que possuem CiteScore maior ou igual a 0.50, no ano de 2017, o que totalizou o universo de 15 títulos, os quais atendiam a esse critério. (Quadro 1).

Quadro 1 – Periódicos da área da Ciência da Informação com CiteScore ≥ 0.50.

Periódico	CiteScore
Scientific Data	6.08
Journal of Cheminformatics	3.98
College and Research Libraries	1.70
Library and Information Science Research	1.70
Communications in Information Literacy	1.44
Journal of the Medical Library Association - JMLA	1.14
Informing Science	1.04
Em busca	0.88
D-Lib Magazine	0.86
Information Research	0.84
Revista Espanola de Documentacion Cientifica	0.83
LIBER Quarterly	0.73
Journal of Information Literacy	0.68
Journal of Information and Organizational Sciences	0.55
Language Documentation and Conservation	0.50

Fonte: Dados da Pesquisa, 2018.

Posteriormente à identificação dos periódicos, partimos para uma visita às *homepages* dos mesmos, em busca de informações sobre a política de compartilhamento de dados que adotam. Ressalta-se que foram visitadas as abas sobre "Política" e "Guia para autores" de todos esses títulos em busca de informações sobre as políticas ou diretrizes de apoio e incentivo ao compartilhamento dos dados científicos.

4 RESULTADOS

Mediante as informações obtidas, elaboramos o Quadro 2, que relaciona os periódicos que apresentam uma política de compartilhamento de dados - PCD de uma maneira geral, e como se dá esse compartilhamento, se de maneira obrigatória – o compartilhamento de dados é inerente ao processo de submissão - ou opcional – o compartilhamento de dados fica a critério do pesquisador, e o periódico apresenta diretrizes para determinada prática.

Quadro 2 – Política de compartilhamento de dados.

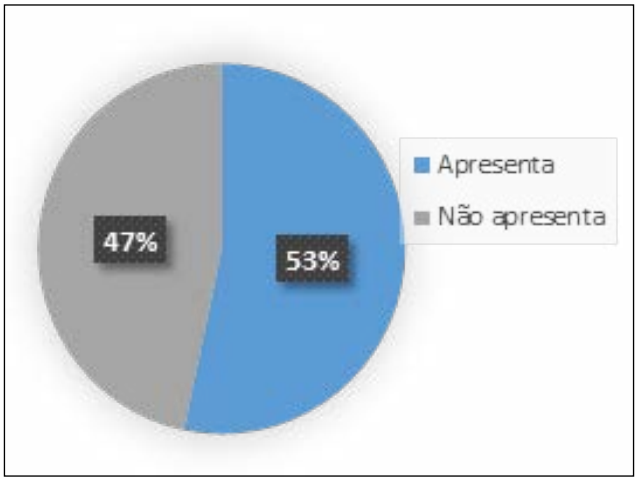
Periódico	PCD	Compartilhamento/Disponibilização
Scientific Data	X	Obrigatório
Journal of Cheminformatics	X	Obrigatório
College and Research Libraries	-	-
Library and Information Science Research	X	Flexível
Communications in Information Literacy	X	Flexível
Journal of the Medical Library Association - JMLA	X	Flexível
Informing Science	X	Flexível
Information Technology and Libraries	-	-
D-Lib Magazine	-	-
Information Research	-	-
Revista Espanola de Documentacion Cientifica	-	-
LIBER Quarterly	X	Flexível
Journal of Information Literacy	-	-
Journal of Information and Organizational Sciences	X	Flexível
Language Documentation and Conservation	-	-

Fonte: Dados da Pesquisa, 2018.

Conforme exposto no Quadro 2, notamos a existência de periódicos que apresentam a política de compartilhamento de dados, seja essa política obrigatória, ou apenas informa aos autores como proceder com essa prática, de forma flexível, deixando que os mesmos escolham como conduzir os seus dados. Descreveremos resumidamente de que forma ocorrem ambas as políticas, obrigatória ou flexível.

Observamos que, dentre os 15 periódicos selecionados, oito apresentam políticas e/ou diretrizes sobre o compartilhamento de dados, representando 53%, enquanto que nos outros sete (47%) não foram encontradas informações acerca do assunto. (Gráfico 1).

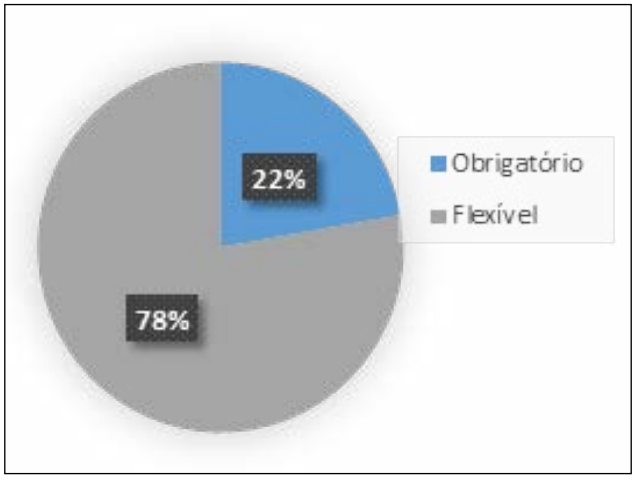
Gráfico 1 - Política de compartilhamento de dados.



Fonte: Dados da Pesquisa, 2018.

Dos oito periódicos que apresentam política ou diretrizes para o compartilhamento de dados, duas exigem que os autores compartilhem os dados oriundos de suas investigações, o que corresponde a dois periódicos (22%), e outras seis (seis periódicos, 78%), sugerem em suas diretrizes formas de compartilhamento dos dados científicos, conforme mostra o Gráfico 2.

Gráfico 2 – Adoção da política de compartilhamento de dados.



Fonte: Dados da Pesquisa, 2018.

No que concerne aos dois periódicos que exigem a prática de compartilhamento de dados, observamos que o periódico Scientific Data, hospedado na Nature.com, lidera no *ranking* do CiteScore da Scopus de 2017. Este periódico apresenta como condição para publicação, a disponibilização de materiais, dados, códigos e/ou protocolos associados à pesquisa (NATURE, 2018).

Nessa mesma perspectiva, o periódico Journal of Cheminformatics, também argumenta que a submissão de um artigo implica que os materiais descritos, incluindo os dados brutos, *softwares* e algoritmos relevantes para chegar a conclusão da pesquisa, deverão estar disponíveis (JOURNAL OF CHEMINFORMATICS, 2018.

No que concerne ao incentivo à prática do compartilhamento/disponibilização dos dados de maneira flexível, o periódico Library and Information Science Research, da editora Elsevier, apresenta diretrizes que incentivam a prática do compartilhamento de dados, além de encorajar o depósito dos dados da pesquisa em um repositório de dados relevante e citar/vincular este conjunto de dados em seu artigo (ELSEVIER, 2018).

Por sua vez, o periódico Communications in Information Literacy informa que os dados coletados com a pesquisa podem ser enviados como arquivo suplementar, ficando a critério do pesquisador incluí-los ou não (COMMUNICATIONS IN INFORMATION LITERACY, 2018).

Outro periódico, que também deixa a critério do pesquisador o compartilhamento/disponibilização dos dados é o Journal of the Medical Library Association, o qual recomenda que, ao serem solicitados, os pesquisadores disponibilizem os dados utilizados em sua investigação, além de recomendar que os pesquisadores mantenham esses dados por um período de cinco anos (JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION, 2018)

O Informing Science também deixa a critério do pesquisador o compartilhamento/disponibilização dos dados, mas incentiva a prática solicitando que disponibilize os dados aos leitores por meio de um *link* (inserido no artigo) em que os dados estejam armazenados (INFORMING SCIENCE, 2018).

O periódico LIBER Quarterly argumenta que, quando apropriado, e a critério do pesquisador, os dados devem ser enviados juntamente com o trabalho, e devem ser armazenados no repositório de dados Dataverse Harvard. Recomenda, ainda, que os pesquisadores atribuam um identificador persistente ao conjunto de dados apropriados ao entendimento da pesquisa e o forneçam na elaboração, submissão e publicação do artigo no periódico (LIBER QUARTERLY, 2018).

Por fim, o Journal of Information and Organizational Sciences destaca que a comissão editorial faz uso dos dados para orientar seu trabalho na publicação e aprimoramento do periódico, deixando a critério do pesquisador o seu compartilhamento/disponibilização, de forma que, se o fizerem, o periódico garante a privacidade dos dados, em termos de violação, acesso, portabilidade, entre outros (JOURNAL OF INFORMATION AND ORGANIZATIONAL SCIENCES, 2018).

Percebe-se que, dentre os periódicos selecionados, 53% apresentam incentivo à prática de compartilhamento de dados, seja de forma obrigatória, ou flexível, deixando que o próprio pesquisador decida o destino deseus dados.

4 CONSIDERAÇÕES FINAIS

A política de compartilhamento de dados há de ser amplamente utilizada por periódicos das mais diversas áreas do conhecimento, visando a colaboração entre pesquisadores, o desenvolvimento da comunicação científica, e o progresso da ciência.

Como visto, a prática do compartilhamento de dados nos periódicos de Ciência da Informação, encontra-se em fase inicial, porém com perspectivas de se consolidar, o que vem a corroborar com o fato de que a comunicação científica se encontra diante de um novo paradigma, ou seja: o compartilhamento/disponibilização dos dados de pesquisa, conforme preconiza a e-Science.

A pesquisa revelou que, dos 15 títulos analisados, oito praticam o compartilhamento/disponibilização dos dados, seja de forma obrigatória ou opcional. Dessa forma, dois periódicos exigem que os autores depositem em repositórios públicos o conjunto de dados utilizados na investigação. Coincidentemente, esses periódicos foram os mais procurados e citados no ano de 2017 pela comunidade científica da Ciência da Informação e lideram o *ranking* do CiteScore.

Os outros sete periódicos incentivam os pesquisadores a compartilharem seus dados possibilitando seu uso e reuso pela comunidade científica.

Concluímos citando Mueller e Passos (2000, p. 20) que as TDIC possibilitam "[...] um mundo mais justo, em que o acesso à informação [dados] fosse estendido a todos os cantos da Terra, e onde o isolamento dos cientistas do Terceiro Mundo fosse amenizada". Dessa forma podemos afirmar que as TDIC proporcionaram transformações na comunicação científica e que essas transformações continuam em curso e as potencialidades tecnológicas desafiam a transição.

REFERÊNCIAS

ANJOS, R. L. **Desvendando as percepções e práticas adotadas pelos pesquisadores dos Programas de Pós-Graduação em Ecologia do Nordeste brasileiro na gestão dos dados científicos**. 2016. 63 p. Trabalho de Conclusão de Curso (Biblioteconomia) – Universidade Federal da Paraíba, João Pessoa, 2016.

COMMUNICATIONS IN INFORMATION LITERACY. **Author Guidelines.** Online. Disponível em: < https://pdxscholar.library.pdx.edu/comminfolit/policies.html>. Acesso em: 25 set. 2018.

ELSEVIER. **Research Data Guidelines**. Online. Disponível em: < https://www.elsevier.com/authors/author-services/research-data/data-guidelines>. Acesso em: 25 set. 2018.

GRAY, J. **eScience**: a transformed scientific method. Palestra apresentada no Conselho Nacional de Pesquisa dos Estados Unidos (NRC-CSTB). Mountain View, Califórnia, 11 jan. 2007. Disponível em: <http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt>. Acesso em 25 set. 2018.

INFORMING SCIENCE. **Article Submission Guidelines.** Online. Disponível em: < https://www.informingscience.org/Journals/InformingSciJ/ArticleSubmission>. Acesso em: 25 set. 2018.

JOURNAL OF INFORMATION AND ORGANIZATIONAL SCIENCES. **Submissions**. Online. Disponível em: < https://jios.foi.hr/index.php/jios/about/submissions#authorGuidelines>. Acesso em: 25 set. 2018.

JOURNAL OF THE MEDICAL LIBRARY ASSOCIATION - JMLA. **Author guidelines**: data retention. Online. Disponível em: < http://jmla.mlanet.org/ojs/jmla/about/submissions#authorGuidelines>. Acesso em: 25 set. 2018.

LIBER QUARTERLY. **Submissions**: author guidelines. Online. Disponível em: <https://www.liber-quarterly.eu/about/submissions/>. Acesso em: 25 set. 2018.

MUELLER, S. P. M.; PASSOS, E. J. L. As questões da comunicação científica e a ciência da informação. In: MUELLER, S. P. M.; PASSOS, E. J. L. (Orgs.). **Comunicação científica**. Brasília: Departamento de Ciências da Informação e Documentação da UnB, 2000, 144p.

NASSI-CALÒ, L. Movimento Open Data se consolida internacionalmente. **SciELO em Perspectiva**, 2014. Disponível em: <http://blog.scielo.org/blog/2014/07/14/movimento-open-data-se-consolida-internacionalmente/>. Acesso em: 25 set. 2018.

NATURE. **Authors e referees**: Availability of data, material and methods. Online. Disponível em: <http://www.nature.com/authors/policies/availability.html>. Acesso em 25 set. 2018.

SAYÃO, L. F.; SALES, L. F. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde,** Rio de Janeiro. v. 8, n. 2, p. 76-92, 2014. Disponível em: <http://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/611/1252>. Acesso em: 26 set. 2018.

TARGINO, M. G. A região geográfica como fator interveniente na produção de artigos de periódicos científicos. In: MUELLER, S. P. M.; PASSOS, E. J. L. (Orgs.). **Comunicação cientifica**. Brasília: Departamento de Ciência da Informação da Universidade de Brasília, 2000. p. 51-72.

PORNOGRAFIA INFANTIL:
análise de manuais de pedofilia

CHILD PORNOGRAPHY:
Pedofilia Manuals Analysis

Marcelo da Silva Morreira[1], Enrique Muriel Torrado[(2)]
(1) PGCIN/UFSC, Florianópolis/SC, marcelomoreira3000@gmail.com.
(2) PGCIN/UFSC, Florianópolis/SC, enrique.muriel@ufsc.br.

Resumo:
A transformação da sociedade, associada ao desenvolvimento tecnológico e da comunicação, provocou grandes mudanças nas inter-relações sociais. Novos paradigmas, como a Sociedade da Informação, surgiram com a universalização do acesso à informação e da propagação crescente do fluxo de dados, contudo a criminalidade não está alheia ao fenômeno das transformações da sociedade. O surgimento da Internet facilitou que criminosos pudessem propagar práticas ilícitas mais facilmente. Crimes relacionados à violência sexual de crianças e adolescentes ganharam um potencial lesivo muito maior com o advento da rede mundial de computadores, pois pedófilos passaram a compartilhar na *Deep Web* manuais e guias para ensinar práticas e técnicas de violência sexual infantil. Nesse contexto, a presente pesquisa tenta produzir um conjunto de diretrizes e princípios de prevenção do crime de Pornografia Infantil, através da análise de manuais de pedofilia disponibilizados na *Deep Web*. Esta pesquisa caracteriza-se como exploratória e qualitativa, sendo explorada através da análise e categorização de manuais de pedofilia; das percepções de policiais no enfrentamento ao problema; e, das percepções de profissionais da área psicossocial no acolhimento das vítimas. Dessa forma, busca-se fornecer um rol de diretrizes de prevenção que possam subsidiar ações de Estado na promoção de políticas públicas de combate à Pornografia Infantil.
Palavras-chave: Pornografia Infantil. Sociedade da Informação. Prevenção.

Abstract:
The transformation of society, associated with technological development and communication, has brought about great changes in social interrelationships. New paradigms, such as the Information Society, have emerged with the universalization of access to information and the growing spread of data flow, yet crime is not alien to the phenomenon of the transformations of society. The emergence of the Internet has made it easier for criminals to propagate illicit practices more easily. Crimes related to sexual violence against children and adolescents have gained much greater potential with the advent of the global computer network, as pedophiles began sharing Deep Web manuals and guides to teach practices and techniques of child sexual violence. In this context, the present research tries to produce a set of guidelines and principles of prevention of the crime of Child Pornography, through the analysis of pedophilia manuals available in the Deep Web. This research is characterized as exploratory and qualitative, being explored through the analysis and categorization of pedophilia manuals; perceptions of cops in coping with the problem; and of the perceptions of psychosocial professionals in the reception of the victims. In this way, we seek to provide a set of prevention guidelines that can support State actions in the promotion of public policies to combat Child Pornography.
Keywords: Child Pornography. Information Society. Prevention.

1 INTRODUÇÃO

O paradigma da Sociedade da Informação ou Sociedade em Rede demonstra que as atuais transformações sociais são permeadas pelo uso da Tecnologia da Informação e Comunicação – TIC. O advento da Internet afetou as relações sociais, as técnicas de produção, a cultura e o próprio conceito de poder, promovendo transformações nas experiências humanas e na própria estrutura social. Todavia, o uso a rede mundial de computadores também possibilitou a propagação de ações criminosas.

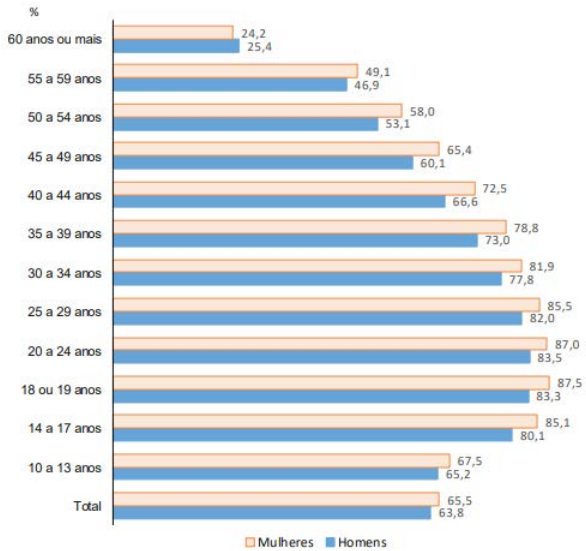
Especialmente, os crimes que envolvem violência sexual contra menores de idade passaram a ter um potencial lesivo ainda maior como o uso crescente do "ambiente virtual". O grande desafio da sociedade em geral, por meio dos órgãos de segurança pública e de entidades de apoio, é buscar medidas efetivas de proteção de suas crianças e adolescentes a fim de combater tais crimes, os quais causam consequências nefastas às vítimas desses abusos.

1.1 Referencial teórico:

Dados da Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), no quarto trimestre de 2016, revelaram que: *i)* 69,3% dos brasileiros têm acesso à rede mundial de computadores; *ii)* em 97,2% dos domicílios em que havia acesso à Internet, o equipamento mais usado era o telefone celular; e, *iii)* 94,2% das pessoas de 10 anos ou mais de idade, usam a Internet com a finalidade de enviar ou receber mensagens de texto, voz ou imagens por aplicativos diferentes de e-mail. (BRASIL, 2016).

Outra informação relevante, apurada na PNAD – Contínua, é o percentual de menores de idade que tem acesso à Internet. No grupo de 10 a 13 anos de idade, as meninas apresentam percentual 67,5%, superior aos meninos (65,2%); e, no grupo de 14 a 17 anos de idade, o percentual das meninas sobe para 85,1% e os meninos para 80,1%. Conforme mostrado no Gráfico 1:

Gráfico 1: Percentual de pessoas que utilizam a Internet – Brasil – 4º trimestre de 2016



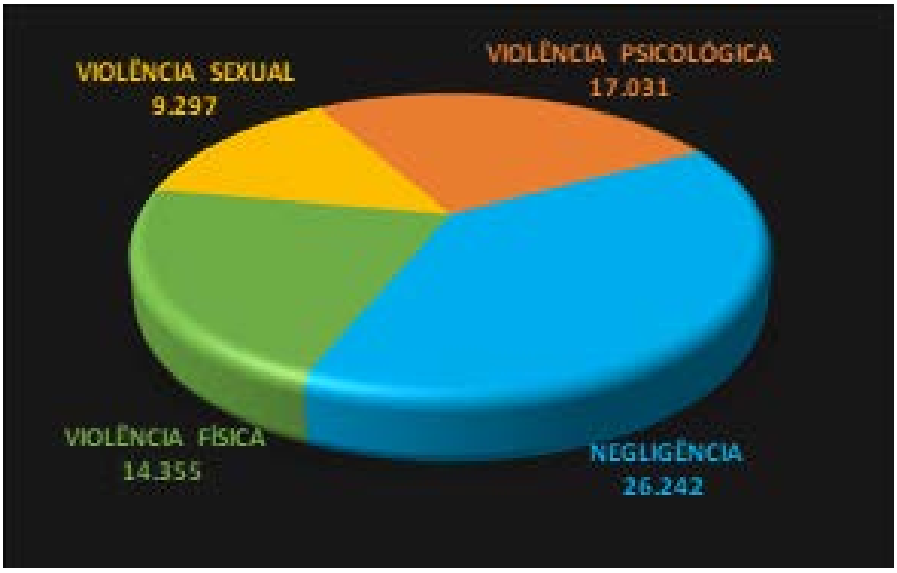
Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional por Amostra de Domicílios Contínua 2016.

De fato, com a evolução da sociedade e o desenvolvimento das tecnologias e das comunicações, a rede mundial de computadores passou a ser recorrida como meio de sociabilidade e de entretenimento, entre crianças e jovens que buscam as redes sociais para passar o tempo e fazer amizades (FERREIRA; SOBRINHO, 2012).

Por outro lado, criminosos usam a Internet com o objetivo de realizar tentativas de aliciamento de possíveis vítimas (crianças e adolescentes) de violência sexual, justamente pelas facilidades tecnológicas oferecidas (CAPPELLARI, 2018) e pela dificuldade de serem identificados.

Em 2018, apenas no período de janeiro a junho, o Ministério dos Direitos Humanos da Presidência da República, por meio do Disk 100, recebeu 9.297 denúncias de violações sexuais contra menores de idade. Nesse levantamento, a violência sexual é a quarta violação mais denunciada, atrás apenas da negligência (26.242 casos), da violência psicológica (17.031 casos) e da violência física (14.355) (BRASIL, 2018). O Gráfico 2 retrata esta realidade.

Gráfico 2: Principais Violações Sexuais contra menores de idade (janeiro a junho de 2018)



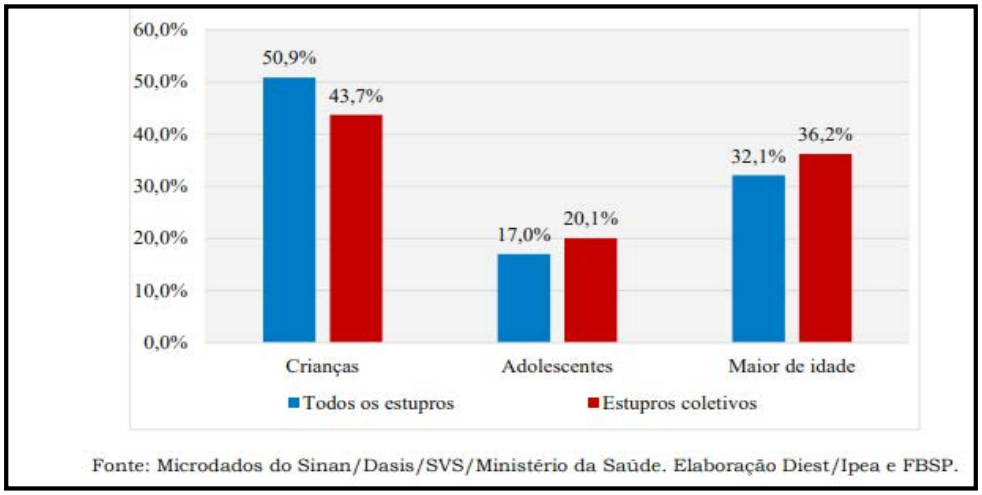
Fonte: Elaborado pelo autor (Dados Disk 100, 2018)

Dessas denúncias de violações sexuais contra crianças e adolescentes:

- 68% estão relacionadas a situações de abuso sexual;
- 15% a exploração sexual; e,
- 14% a pornografia infantil.

Para aprofundar neste grave problema, o Atlas de Violência 2018, produzido pelo Instituto de Pesquisa Econômica Aplicada (Ipea) e pelo Fórum Brasileiro de Segurança Pública (FBSP) revelou que, no ano de 2016, as polícias brasileiras registraram 49.497 casos de estupro e o Sistema Único de Saúde (SUS) notificou 22.918 incidentes desta mesma natureza; sendo que, em quase 68% do total de casos de estupro ocorridos no Brasil, as vítimas eram crianças ou adolescentes; revelando ainda que 63,8% dos casos de estupros coletivos envolviam menores de idade. No gráfico 3, o destaque fica nos casos da participação de crianças e adolescentes nos casos de estupros coletivos.

Gráfico 3: Brasil: distribuição percentual das vítimas de estupro no total de casos e nos estupros coletivos, 2016



Fonte: Microdados do Sinan/Dasis/SVS/Ministério da Saúde. Elaborado Diest/Ipea e FBSP, 2016.

As formas violência sexual infantojuvenil envolvem um cenário complexo de segurança pública, associado aos graves danos físicos e psicológicos acometidos às vítimas, à relação de proximidade dos agressores, e ao local dos fatos que, devido aos avanços tecnológicos, teve, ainda, seu espaço expandido para o ciberespaço (CASTELLS, 1999) por meio da Internet. "[...] diante das novas possibilidades de inter-relação propiciadas pela Internet, poucas vezes a pedofilia esteve tão visível e, ao mesmo tempo, tão difícil de ser controlada" (CAPPELLARI, 2018, p. 1).

Nesse contexto de expansão da criminalidade, pedófilos, na rede mundial de computadores, passaram a disponibilizar *on-line* fotos e vídeos contendo cenas de sexo infatojuvenil, tanto com o intuito de perpetrar sua satisfação pessoal, bem como para utilizar como "moeda de troca" na aquisição de novos vídeos e imagens em comunidades de pedofilia. Por estes motivos, a Internet se tornou o meio mais usado para a distribuição e compartilhamento deste material ilícito (CAPPELLARI, 2018), através de redes sociais, de redes *peer-to-peer* (P2P), de grupos de *Whatsapp* e da rede profunda, conhecida como *Deep Web*.

Cabe ressaltar que pedófilos têm compartilhado na *Deep Web* manuais de pedofilia a fim de divulgar práticas e técnicas de abuso sexual de menores de idade. Tais guias têm como objetivo ensinar:

- Sobre a sexualidade infantil;
- Sobre os riscos envolvidos;
- Quando começar, idade ideal;
- Onde encontrar uma criança (filhos próprios, família, mães com filhos, creches, escolas, na rua;
- Como escolher bem das possíveis vítimas;
- Como abordar ou se aproximar de crianças ou adolescentes;
- Como conquistar a mãe para ter acesso à vítima, menor de idade;
- Como conquistar a confiança da vítima;
- Como concretizar o abuso sem deixar marcas;
- Como fazer a criança te dar prazer; e,
- Como se prevenir da ação policial.

Na "Era da Informação", referida por Burke (2003), a informação tem valor para a sociedade e promove o seu desenvolvimento. Entretanto, a divulgação de manuais para ensinar a praticar o mal vai de encontro ao conceito de Sociedade da Informação (CASTELLS, 1999) e se contrapõe a todos os valores sociais.

O presente estudo busca, através da análise de manuais de pedofilia, de entrevistas com policiais que atuam no enfrentamento do problema e de entrevistas de profissionais da área psicossocial que atual no acolhimento e tratamento das vítimas, propor uma estrutura de diretrizes e princípios de prevenção que ajudem a sociedade no combate à pornografia infantil, servindo de matéria-prima para a elaboração de campanhas publicitárias, palestras educacionais e ações de conscientização.

A questão central a ser respondida é: Como podemos combater e previr a pornografia infantil no contexto da Sociedade de Informação? Com o intuito de responder esta indagação, os seguintes objetivos específicos foram elaborados: a) Analisar as relações da Sociedade da Informação com o contexto do crime de pornografia infantil; b) Identificar e caracterizar os manuais e guias de pedofilia disponibilizados na *Deep Web*; c) Coletar as percepções de policiais que trabalham no combate aos crimes de violência sexual infantojuvenil; d) Verificar e analisar as percepções de profissionais da área de assistência psicossocial que atuam no acolhimento e tratamento de menores vítimas de abusos sexuais, e; e) Identificar os fluxos institucionais dos órgãos públicos que atuam na área de segurança pública e de assistência psicossocial.

2 OBJETIVOS

2.1 Objetivo Geral:

Produzir um conjunto de diretrizes e princípios de prevenção do crime de Pornografia Infantil através da análise de guias e manuais de pedofilia disponibilizados na *Deep Web*.

2.2 Objetivos Específicos:

A Tabela 1 lista os objetivos específicos do presente estudo:

Tabela 1: Objetivos Específicos.

Objetivos Específicos
a) Analisar as relações da Sociedade da Informação com o contexto do crime de pornografia infantil.
b) Identificar e caracterizar os manuais e guias de pedofilia disponibilizados na *Deep Web*.
c) Coletar as percepções de policiais que trabalham no combate aos crimes de violência sexual infantojuvenil.
d) Verificar e analisar as percepções de profissionais da área de assistência psicossocial que atuam no acolhimento e tratamento de menores vítimas de abusos sexuais.
e) Identificar os fluxos institucionais dos órgãos públicos que atuam na área de segurança pública e de assistência psicossocial.

Fonte: Elaborado pelo autor, 2018.

3 PROCEDIMENTOS METODOLÓGICOS

Neste estudo serão analisados 04 manuais de pedofilia e serão entrevistados profissionais da área de segurança e assistência psicossocial envolvidos no contexto do crime de Pornografia Infantil no intuito de propor uma estrutura de diretrizes e princípios que possam dar apoio à prevenção deste tipo de ilícito.

3.1 Caracterização da Pesquisa

A abordagem do problema é qualitativa, em virtude da busca da compreensão dos elementos associados ao contexto da prevenção da Pornografia Infantil. É um estudo de caráter exploratória-descritivo uma vez que que "pressupõem (sic) que o investigador tenha um conhecimento aprofundado a respeito dos fenômenos e problemas que está estudando" (KÖCHE, 2010, p.126). Dessa forma, a meta é analisar a questão da prevenção a partir de diferentes perspectivas, por meio do emprego de entrevistas com questionamentos semiestruturados (ROSA; ARNOLDI, 2006).

A pesquisa é caracterizada quanto aos procedimentos metodológicos como documental, pois "vale-se de materiais que não receberam ainda um tratamento analítico." (GIL, 2008, p.51), como os guias de pedofilia disponibilizados na *Deep Web*.

A tabela 2 detalha os procedimentos metodológicos:

Tabela 2: Procedimentos Metodológicos.

Procedimentos Metodológicos
a_1) Levantamento da legislação pátria vigente relacionada ao crime de pornografia infantil.
a_2) Levantamento bibliográfico de artigos e publicações científicas relacionadas a interseção da Ciência da Informação, pornografia infantil e prevenção.
b) Análise de conteúdo das principais dicas recomendações compartilhadas por pedófilos.
c) Realizar de entrevistas semiestruturadas com policiais que atuam no combate ao crime de pornografia infantil.
d) Realizar de entrevistas semiestruturadas com psicólogos e assistentes sociais que atuam com menores vítimas de violência sexual.
e) Levantamento de protocolos, fluxos e rotinas de trabalho dos órgãos envolvidos no enfrentamento do problema.

Fonte: Elaborado pelo autor, 2018.

Por fim, quanto ao referencial teórico, é importante salientar que neste estudo são apresentados dados, gráficos e análises no intuito de tentar demonstrar a dimensão do problema da violência sexual infantil no Brasil. Entretanto, a presente pesquisa tem como eixo central a extração de conteúdo de manuais de pedofilia disponibilizados na *Deep Web* com o intuito de fomentar a prevenção, dessa forma será produzida a partir de um levantamento bibliográfico (LAKATOS, MARCONI, 2010), através livros e artigos de periódicos que tenham relação com a Ciência da Informação, crimes de violência sexual infantil e prevenção.

A pesquisa da fundamentação bibliográfica será realizada nas bases de dados: *Library and Information Science Abstracts* (LISA), SUCUPIRA, SCOPUS e *Web of Science* (WoS), tendo em vista a grande quantidade de períodicos nestas bases e o seu reconhecimento internacional.

4 RESULTADOS

A iminente pesquisa encontra-se numa etapa inicial, tendo sido realizado o levantamento da legislação pátria relacionada à questão da proteção da criança e do adolescente, bem como da legislação associada aos crimes de violência sexual de menores de idade.

Foi realizada a também a Revisão Sistemática da Literatura relacionada à relação da Ciência da Informação, Pornografia Infantil e prevenção.

Ao longo da pesquisa, serão analisados quatro manuais de pedofilia intitulados:

- "How to practice child love";
- "The Pedophile's Handbook";
- "Pedo playground";
- "DDLG Guide".

A fase atual é a análise e categorização dos manuais de pedofilia disponíveis na *Deep Web*. Tendo sido iniciado a tradução e categorização do primeiro dos quatro manuais de pedofilia disponibilizados na *Deep Web*.

A tabela 3 apresenta os resultados realizados e esperados relacionando-os aos objetivos específicos e procedimentos metodológicos:

Tabela 3: Resultados.

Resultados
a_1) Resultado Esperado: Fundamentação legal sobre o crime de Pornografia Infantil. **a_1) Resultado Realizado:** I. Arts. 109, 144, 227 da Constituição da República Federativa do Brasil II. Convenção sobre os Direitos da Criança adotada pela Assembleia Geral das Nações Unidas III. Lei nº 8069/90 (Estatuto da Criança e do Adolescente - ECA) IV. Lei nº 11829/2008 (alterou o ECA) V. Decreto 99.710/90 - Convenção sobre os Direitos da Criança VI. STJ, RECURSO ESPECIAL Nº 1.543.267 – SC (03/12/2015) VII. STF, RECURSO EXTRAORDINÁRIO 628.624 (Minas Gerais);
a_2) Resultado Esperado: Fundamentação teórica sobre as interseções da Sociedade da Informação com o crime de Pornografia Infantil. **a_2) Resultado Realizado:** I. Revisão Sistemática da Literatura.
b) Resultado Esperado: Categorização das dicas e recomendações compartilhadas em manuais de pedofilia na *Deep Web*. **b) Resultado Realizado:** Em andamento.
c) Resultado Esperado: Identificação das boas práticas de combate ao crime de Pornografia Infantil no contexto policial; coleta de experiências e captação de dicas de prevenção.
d) Resultado Esperado: Identificação dos aspectos relacionados às vítimas de violência sexual infanto-juvenil, observadas por profissionais na área de atendimento e de acolhimento.
e) Resultado Esperado: Identificação de protocolos, fluxos e rotinas no encaminhamento de denúncias de Pornografia Infantil. Levantamento das etapas de acolhimento e tratamento de vítimas pelos órgãos públicos.

Fonte: Elaborado pelo autor, 2018.

4 CONSIDERAÇÕES FINAIS

Ao final, busca-se que, à luz da Ciência da Informação, o resultado desta pesquisa possa efetivamente contribuir para a redução dos casos de violação sexual de crianças e adolescentes, enumerando um conjunto de diretrizes e princípios de prevenção que ajudem a sociedade no combate à pornografia infantil, servindo de matéria-prima para a elaboração de campanhas publicitárias, palestras educacionais e ações de conscientização.

Espera-se que o conjunto de diretrizes e princípios de prevenção produzidos por este estudo possam compor manuais de orientação a pais, professores, policiais e sociedade em geral no intuito de permitir maior proteção de crianças e adolescentes em relação aos crimes de violência sexual infantil.

REFERÊNCIAS

BRASIL. Instituto Brasileiro de Geografia e Estatística. **Pesquisa Nacional por Amostra de Domicílios Contínua – PNAD Contínua.** 2016. Disponível em: <ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Anual/Acesso_Internet_Televisao_e_Posse_Telefone_Movel_2016/Analise_dos_Resultados.pdf>. Acesso em: 10 set. 2018.

BRASIL. Instituto de Pesquisa Econômica Aplicada. **Atlas da Violência.** 2018. Disponível em:<http://www.ipea.gov.br/portal/images/stories/PDFs/relatorio_institucional/180604_atlas_da_violencia_2018.pdf>. Acesso em: 19 set. 2018.

BURKE, Peter. **Uma história social do conhecimento - I: de Gutenberg a Diderot.** Trad. por Plínio Dentzien. Rio de Janeiro: Zahar, 2003.

CAPPELLARI, M. R. S. V. A pedofilia na pós-modernidade: um problema que ultrapassa a cibercultura. **Em Questão**, v. 11, n. 1, p. 67-82, 2005. Disponível em: <http://www.brapci.inf.br/index.php/article/view/0000003734/fcd6ae2b54c19896ab4cce32c6662b37>. Acesso em: 14 set. 2018.

CASTELLS, Manuel. **A Sociedade em Rede:** A Era da Informação: Economia, Sociedade e Cultura. 8. ed. São Paulo: Paz e Terra, 1999. 44 p. 1 v. Tradução de: Roneide Venancio Majer.

FERREIRA, Rubens da Silva; SOBRINHO, Telma Socorro da Silva. Bibliotecas públicas e ciberviolência em tempos de uma sociedade em rede: Novos papéis diante de um novo fenômeno. **Revistas de La Fahce**: Facultad de Humanidades y Ciencias de la Educación/UNPL, Buenos Aires, Argentina, v. 2, n. 1, 2012. Disponível em: <https://www.palabraclave.fahce.unlp.edu.ar/article/view/PCv-2n1a03>. Acesso em: 14 set. 2018.

GIL, A. C. **Métodos e técnicas de pesquisa social.** 6. ed. São Paulo: Atlas, 2008. 51 p.

KÖCHE, J. C. **Fundamentos da metodologia científica.** 27. ed. Petrópolis: Vozes, 2010. 126 p.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de Metodologia Científica.** 7. ed. São Paulo: Atlas, 2010.

ROSA, M. V. F. P. C.; ARNOLDI, M. A. G. C. **A entrevista na pesquisa qualitativa:** mecanismos para a validação dos resultados. Belo Horizonte: Autêntica Editora, 2006. 112 p.

PRÁTICAS DE GESTÃO DE DADOS:
uma Revisão da Literatura Sobre o Termo *Data Life Cycle*

DATA MANAGEMENT PRACTICES:
a Review of the Term Data Life Cycle

Débora Gomes de Araújo[1]
(1) Universidade Federal da Paraíba (UFPB), Cidade Universitária, s/n - Castelo Branco III, João Pessoa - PB, 58051-085, debora.g.de.araujo@gmail.com.

Resumo:
Este trabalho realiza uma revisão bibliográfica sobre o termo "*data life cycle*", com enfoque na área da Ciência da Informação, tendo as bases de dados Emerald, LISA e LISTA como fontes de pesquisa. Nesse estudo foi utilizada uma abordagem quantitativa, de cunho exploratório e bibliográfico. Foram utilizadas as ferramentas Zotero e QDA Miner na condução da investigação. Quanto aos resultados, ficou evidenciado que a base de dados que mais indexou sobre a temática em questão foi a LISTA no período de 2013 a 2018, porém ainda apresentou uma quantidade pouco expressiva de indexações. O autor que se destacou na quantidade das produções científicas sobre o tema foi Vardigan, enquanto que o periódico que teve o maior número de publicação foi o americano *International Association for Social Science Information Service and Technology Quarterly* e o Reino Unido foi o local que mais publicou. Por meio dos artigos recuperados foi possível identificar a importância da gestão de dados em cada etapa do ciclo de vida de dados. Desta forma, este trabalho contribui para facilitar o entendimento da área pelos pesquisadores que investigam o tema.
Palavras-chaves: *Big data*. Ciência da Informação. Ciclo de vida dos dados. Gestão de dados científicos.

Abstract:
This work review the term "data life cycle" in the literature focused in the field of Information Science. Emerald, LISA and LISTA databases were used as sources of research. The study applied a quantitative approach with exploratory and bibliographic efforts. Zotero and QDA Miner tools were used to support the research. Although a little expressive amount of indexations have been found from 2013 to 2018, the results evidenced LISTA as the most indexed database. In addition, the recovered papers highlighted Vardigan as the most productive author in the field. The results also evidenced the American International Association for Social Science Information Service and Technology Quarterly as the most published periodical and the United Kingdom as the most published place. Through the retrieved papers it was possible to identify the importance of data management in each step of the data life cycle. This work will contribute to understanding the subject by the researchers in the field.
Keywords: Big data. Information Science. Data life cycle. Scientific data management.

1 INTRODUÇÃO

Nos dias atuais, cada vez mais a gestão de dados gerados a partir das pesquisas científicas vem ganhando espaço, devido as necessidades de uso e reuso dos mesmos. O que é facilitado pelas tecnologias contemporâneas. Neste sentido, com a disseminação da tecnologia da informação e comunicação (TIC) em nosso cotidiano, a pesquisa científica também evoluiu, em um contexto relacionado com o uso intenso dos dados, através da obtenção de informações de grandes volumes de dados digitalizados (AYDINOGLU; DOGAN; TASKIN, 2017).

O estudo em questão tem o seu foco nos dados de pesquisa, que de acordo com Patel (2016) estes constituem o centro de qualquer investigação científica, pois as descobertas e conclusões dos estudos são totalmente dependentes deles.

Dados de pesquisa incluem-se no processo dinâmico de investigação científica. Neste cenário, eles podem ser distinguidos por meio de duas funções: dados como materiais (*input*), está relacionada com a fase inicial, momento em que os dados são coletados e analisados e os dados como resultado (*output*), refere-se aos dados que são produzidos no decorrer dos processos, levando a publicações, como finalidade da pesquisa (SCHÖPFEL et al., 2016).

As práticas de pesquisas atuais exigem novas formas de tratamento dos dados por parte dos pesquisadores, de forma a acompanhar as mudanças constantes. Diversas áreas do conhecimento desenvolvem planos de gestão de dados por exigências de agências fomentadoras de pesquisa. O ciclo de vida de dados (CVD) se apresenta nesse cenário como uma ferramenta que pode promover habilidades e conhecimento para o pesquisador conduzir de forma apropriada a sua pesquisa, oferecendo etapas que contemple todo o percurso dos dados, de forma a serem detectáveis e utilizáveis em outros estudos.

Desta forma, é fundamental compreender o CVD, pois além de ser essencial em seu próprio trabalho, contribui com os pesquisadores, ao possibilitar soluções para as barreiras que podem ser encontradas na coleta e análise de dados, assim como na organização de conjuntos de dados e na descoberta de conjuntos externos relevantes (GOBEN; RASZEWSKI, 2015).

Segundo Rice e Southall (2016), um ciclo de vida de pesquisa mapeia a atividade de um pesquisador durante um projeto de pesquisa. Analogamente, um CVD representa o percurso dos dados ou as ações necessárias para que a pesquisa evolua para o estágio seguinte.

Na literatura existem algumas iniciativas que representam os ciclos de vida dos dados. De acordo com Goben e Raszewski (2015), existem vários modelos capazes de descrevê-los, porém eles citam como exemplos mais populares o *Digital Curation Center* (DCC) e a iniciativa do *Data Observation Network for Earth* (DataONE). Outras iniciativas podem ser evidenciadas como o *Data Documentation Initiative* (DDI), apresentada por Vardigan (2013), assim como os modelos destacados por Rice e Southall (2016), da *Joint Information Systems Committee* (Jisc) (2016) e do *Imperial College London Library RDM Workflow*, cujo autor é Barnes (2016), existem ainda diversos modelos de ciclo de vida de dados.

Da realidade da coleta, armazenamento, recuperação e descarte de dados, Santa'Ana (2016) desenvolveu um modelo voltado para a ciência da informação (CI), uma vez que, enfatiza o papel desta área no estudo e na proposta de caminhos para lidar adequadamente com os dados, ao revelar que a área pode trabalhar com um novo enfoque, sendo uma aliada nesse processo de otimização de dados.

2 OBJETIVOS

Diante destas considerações iniciais, o objetivo da presente pesquisa foi realizar uma revisão bibliográfica sobre o termo *data life cycle*, na literatura concentrada na área da Ciência da Informação (CI), nas bases de dados *Emerald eJournals Premier* (Emerald), *Library and Information Science Abstracts* (LISA) e *Library e Information Science & Technology Abstracts* (LISTA). Uma vez que, estas bases registram de formas variadas e focam de forma específica na CI.

3 PROCEDIMENTOS METODOLÓGICOS

O estudo em questão é de cunho quantitativo, exploratório e bibliográfico. Teve como fonte de dados os metadados dos artigos de arquivos abertos encontrados nas bases de dados supracitadas. A fase de coleta compreendeu o período de 01/07/2018 a 23/07/2018. Foram adotados os seguintes critérios de busca: utilizou-se apenas o termo "*data life cycle*", o qual foi coletado entre aspas, o espaço temporal de levantamento foi entre 2013-2018, com a finalidade de identificar o que vem sendo trabalhado sobre o tema nos últimos cinco anos.

Os dados revelaram que a base que mais indexou artigos de arquivos abertos sobre o *data life cycle* no período em questão foi a LISTA, pois através dela foi possível recuperar 35 artigos. Sendo 22 provenientes da Emerald e 14 da LISA. Em 2015 houve um avanço nas produções, o que pôde ser verificado nas três bases. Em 2016 o volume foi significativo. Apresentou uma queda em 2017 na base LISTA, pois não houve indexações. Em 2018 as referidas bases continuam disponibilizando artigos sobre a temática. Desta forma, o tema é alvo de discussão no cenário atual.

Os artigos oriundos da busca bibliográfica foram exportados para a ferramenta ZOTERO, através desta foi possível constatar na opção itens duplicados, os artigos recuperados que se repetiam entre as bases de dados. Através dessa análise, verificou-se que dos 14 artigos encontrados na base de dados LISA, 12 deles estão presentes nas bases de dados Emerald e LISTA, sendo 10 na Emerald e 2 na LISTA. Em seguida, os registros duplicados foram eliminados, passando do total de 71 para 59 artigos.

4 RESULTADOS

A partir do levantamento dos autores que produziram sobre o tema, constatou-se que apenas 13% deles publicaram mais de uma vez sobre o termo *data life cycle*. Enquanto uma quantidade significativa de autores, representada por 87%, discorreu pouco sobre o tema, pois apresentaram apenas um artigo. Infere-se que uma quantidade reduzida de estudiosos produziu um pouco mais sobre a temática, ao passo que um número considerável de pesquisadores está publicando pouco.

A autora que se destacou na quantidade de artigos sobre o tema foi Mary Vardigan, produziu 4 artigos nos períodos de 2013 a 2016, no periódico americano *IASSIST Quarterly*, sendo que a

produção de 2013 foi apenas de sua autoria e as demais envolveram outros pesquisadores, sendo duas delas com Ionescu.

Constatou-se que os autores Borgman, Darch, Sands, Wallis e Traweek produziram dois artigos juntos. Os pesquisadores Schöpfel, Južnič, Prost, Malleret, Češarek e Koler-Povh em conjunto produziram dois artigos e os autores Si, Xing, Zhuang e Hua também tiveram dois trabalhos juntos. Desta forma, a maioria dos estudiosos que apresentaram mais de um artigo tiveram seus trabalhos em parceria com outros que se enquadraram nas mesmas condições, com exceção de Harris que publicou sozinho dois artigos e Li que desenvolveu suas pesquisas com outros pesquisadores que tiveram apenas 1 publicação.

A análise possibilitou levantar os periódicos que apresentaram mais produções no período sobre a temática. Destacamos os que têm acima de uma publicação, conforme o Quadro 1.

Quadro 1: Distribuição dos periódicos por países no período de 2013-2018.

TÍTULO DE PUBLICAÇÃO	NÚMERO	PAÍSES
IASSIST Quarterly	8	Estados Unidos
International Journal on Digital Libraries	3	Alemanha
Information Services & Use	3	Holanda
Journal of Map & Geography Libraries	3	Estados Unidos
Information & Computer Security	2	Inglaterra
Journal of The Association for Information Science and Technology	2	Estados Unidos
Library Hi Tech	2	Inglaterra
New Review of Information Networking	2	Inglaterra
Program: electronic library and information systems	2	Inglaterra
The Electronic Library	2	Inglaterra
The Canadian Journal of Library and Information Practice and Research	2	Canadá
The Grey Journal	2	Holanda

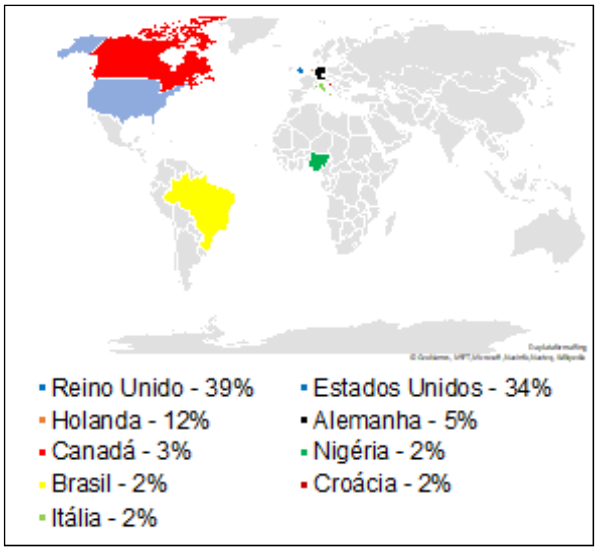
Fonte: Dados da pesquisa, 2018.

Através do levantamento dos Títulos de publicações, foi possível constatar que o maior número de publicações ocorreu no periódico americano *IASSIST*[1] *Quarterly*, ao apresentar 8 produções.

Ao incluir os artigos de outros periódicos que tiveram apenas uma produção, foi possível verificar os países que trataram da temática. O que está expresso na Figura 1.

1 International Association for Social Science Information Service and Technology

Figura 1: Países que publicaram sobre *data life cycle* no período de 2013-2018.



Dados da pesquisa, 2018.

Diante dos dados apresentados, é possível identificar que o *data life cycle* é um tema discutido no cenário nacional e internacional, com uma concentração de publicações em periódicos do continente europeu, representando 60%, o que pode ser resultado de iniciativas como o Horizon 2020 (H2020), que consiste em um programa de pesquisa e inovação que tem por finalidade aperfeiçoar o acesso à informação científica, no que tange aos artigos de pesquisa científica e aos dados de pesquisa. Em um contexto digital de Open Acess (HORIZON, 2018).

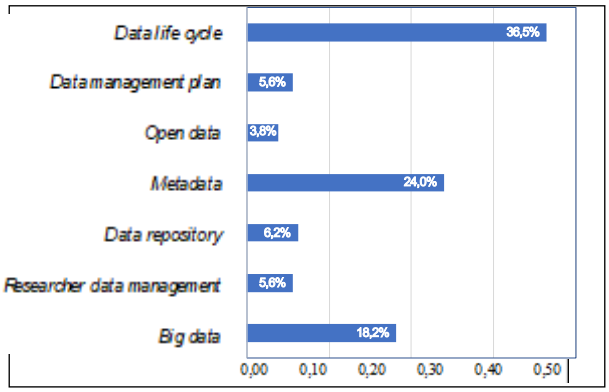
Nesse cenário, observa-se que o Reino Unido se destacou com um percentual de 39%, o que pode remeter a ser um local de referência na criação de ciclos de vida de dados, como por exemplo, o DCC, a iniciativa da Jisc e a do *Imperial College London Library RDM Workflow.* Além dessas estratégias voltadas para lidar corretamente com os dados de pesquisas, Hua et al. (2015) destacam que a biblioteca da Universidade de Cambridge oferece treinamentos para os pesquisadores gerenciarem seus dados.

Posteriormente, aparece o continente americano com 39% das produções, sendo 34% dos Estados Unidos. Diante desta realidade, constata-se a presença de iniciativas que visam a gestão de dados no cenário americano, como a criação dos ciclos de vida de dados DataOne e o DDI. Com uma participação reduzida de 2%, encontramos também uma produção brasileira, que também apresenta um modelo de ciclo de vida de dados com o foco na CI, do autor Sant'Ana (2016).

O estudo revela que apesar do autor com maior número de publicações e o periódico que mais se destacou serem americanos, o local que mais produziu atualmente sobre o data life cycle é o Reino Unido, seguido dos Estados Unidos.

Tomando como base as palavras-chave dos artigos citados pelos autores que publicaram em mais de um artigo, por meio da ferramenta QDA MINER, foi realizada uma pesquisa, a partir da criação de códigos nomeados pelas palavras *data life cycle* (ciclo de vida de dados), big data, cujo termo se refere a volume, variedade, velocidade e veracidade dos dados, segundo Dale (2015), *research data management* (gestão de dados de pesquisa), *data repositor* (repositórios de dados), *metadata* (metadados), open data (acesso aberto) e *data management plan* (plano de gestão de dados), para verificar a frequência desses termos em todos os arquivos recuperados, exceto os que estavam presentes nas referências dos artigos analisados (Gráfico 1).

Gráfico 1: Frequência dos termos encontrados nos artigos sobre *data life cycle* utilizando a ferramenta QDA Miner.

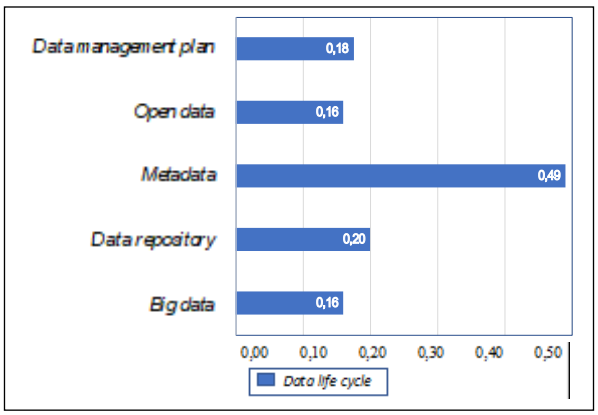


Fonte: Dados da pesquisa, 2018.

O Gráfico 1 confirma que o termo mais frequente nos textos avaliados foi o *data life cycle*, haja vista que os arquivos recuperados nas bases de dados se basearam no referido termo. Os demais termos estão envolvidos com a temática com destaque para *metadata* com 24% e *big data* com 18,2%. Por meio desse levantamento foi possível codificar os textos com as palavras supramencionadas, possibilitando a extração de segmentos relevantes contidos nos textos para o apoio teórico da pesquisa.

O Gráfico 2 a seguir, apresenta a coocorrência dos termos com relação ao *data life cycle*, ou seja, as vezes que as citações dos termos ocorreram simultaneamente, o que revela a proximidade dos mesmos.

Gráfico 2: Proximidade dos termos encontrados nos artigos sobre *data life cycle* utilizando a ferramenta QDA Miner.



Fonte: Dados da pesquisa, 2018

Apesar do *big data* ter sido o termo mais citado nos textos recuperados pelo Gráfico 1, ele juntamente com o *open data* são os mais distantes do termo *data life cycle* na análise de proximidade (Gráfico 2) e o *Researcher data management* não apresentou coocorrência. O termo *metadata* foi recuperado como o mais próximo, seguido do *data repository*. Essa análise se mostra importante quando na recuperação dos termos coocorrentes, o que ajuda na identificação de termos simultâneos.

5 CONSIDERAÇÕES FINAIS

Através da pesquisa realizada foi possível fazer uma revisão da literatura sobre o termo *data life cycle*, nas bases de dados Emerald, LISA e LISTA. Evidenciou-se que a temática ainda é emergente na área da Ciência da Informação, uma vez que, a quantidade de produções não foi expressiva no período analisado. No entanto, a *IASSIST QUARTERLY* se destacou pela quantidade de publicações na área da CI, denotando alinhamento com os interesses dos profissionais da área sobre a temática, mesmo que ainda seja pouco explorada.

O termo é discutido no cenário internacional, com destaque para o Reino Unido, visto o maior de número de produções. Mary Vardigan, que mais produziu no período estudado, trabalha em projetos relacionados à administração de dados, tendo atuado como diretora do DDI, que como dito anteriormente, representa um ciclo de vida dos dados, podendo ser considerada uma referência no assunto.

As ferramentas Zotero e QDA Miner ofereceram um apoio na condução da pesquisa. Através da segunda foi possível extrair dos textos citações relevantes para o trabalho, facilitando o processo de pesquisa, em que a partir dos segmentos selecionados ficou claro que o processo de gerenciamento de dados científicos está diretamente ligado com o ciclo de vida dos dados, sendo tal conexão algo fundamental para possibilitar o compartilhamento dos dados, uma vez que, boas práticas de gerenciamento são necessárias em cada fase do ciclo de vida dos dados.

Recomendamos como sugestão de trabalhos futuros, o desenvolvimento de pesquisas que, além de fazer um levantamento dos autores que mais publicaram na área, mostrem a quantidade de citações que têm recebido. Sugerimos ainda que estudos sejam realizados para identificar os ciclos de vida de dados existentes no âmbito científico, empresarial e governamental.

REFERÊNCIAS

AYDINOGLU, A. U.; DOGAN, G.; TASKIN, Z. Research data management in Turkey: perceptions and practices. **Library Hi Tech**, v. 35, n. 2, p. 271–289, 27 abr. 2017. Disponível em< https://www.emeraldinsight.com/doi/abs/10.1108/LHT-11-2016-0134> Acesso em 17 de jul. 2018.

BARNES, A. Imperial Colegge London Library RDM Workflow. Zenodo. 2016. Disponível em:< https://zenodo.org/record/54000#.W6FzFmhKjIU>. Acesso em 18 set. 2018.

DALE, K. L. RIM's Role in Harnessing the Power of Big Data. **Information Management Journal**, v. 49, n. 4, p. 29–32, 7 ago. 2015.

DATAONE. **Primer on Data Management**: what you always wanted know. Disponível em: <https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf>. Acesso em: 20 jun.2018.

DCC Digital Curation Centre. **Curation Lifecycle Model**. Disponível em: <http://www.dcc.ac.uk/resources/curation-lifecycle-model> Acesso em: 20 jun. 2018.

GOBEN, A.; RASZEWSKI, R. The data life cycle applied to our own data. **Journal of the Medical Library Association**, v. 103, n. 1, p. 40–44, jan. 2015. DOI: http://dx.doi.org/10.3163/1536-5050.103.1.008. Disponível em< https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4279933/> Acesso em 17 de jul. 2018.

H2020. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Disponível em:<http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf>. Acesso em 8 ago. 2018.

HUA, X. et al. Investigation and analysis of research data services in university libraries. **The Electronic Library**, v. 33, n. 3, p. 417–449, 27 maio 2015.

Joint Information Systems Committee.2016. Disponível em< https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data> Acesso em: 28 set. 2018.

PATEL, D. Research data management: a conceptual framework. **Library Review**, v. 65, n. 4/5, p. 226–241, maio 2016.

QDA MINER. Disponível em< https://provalisresearch.com/products/qualitative-data-analysis-software/> Acesso em: 25 jul. 2018.

RICE, R.; SOUTHALL, J. **The Data Librarian´s handbook.** Publisher by Facet Publishing. London, 2016.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, v. 21, n. 2, p. 116, 20 dez. 2016.

SCHÖPFEL, J. et al. Dissertations and Data. **Grey Journal (TGJ)**, v. 12, n. 3, p. 126–148, set. 2016.

VARDIGAN, M. The DDI Matures: 1997 to the Present. **IASSIST Quarterly**, v. 37, n. 1–4, p. 45–50, mar. 2013.

ZOTERO. Disponível em< https://www.zotero.org/> Acesso em 27 jul. 2018.

PREVISÕES DE AUSÊNCIA EM ATENDIMENTO ESPECIALIZADO EM UNIDADES DE SAÚDE PÚBLICA DE FLORIANÓPOLIS

ABSENCE FORECAST IN SPECIALIZED CARE IN PUBLIC HEALTH UNITS OF FLORIANÓPOLIS

Milena Teixeira[1], Leandro Pereira Garcia[(2,3)], Ingo Ramos[(1)], Lucas Alexandre Pedebôs[(3)], Douglas Macedo[(1)], Gustavo Medeiros de Araujo[(1)]

(1) Universidade Federal de Santa Catarina, R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900, milena.correia@grad.ufsc.br,ingoramos12@gmail.com douglas.macedo@ufsc.br,gustavo.araujo@ufsc.br.

(2) Universidade do Sul de Santa Catarina, Avenida Pedra Branca, 25, Cidade Universitária Pedra Branca, CEP 88137-272, Palhoça, SC lpgarcia18@gmail.com

(3) Secretaria Municipal de Saúde de Florianópolis, Avenida Professor Henrique da Silva Fontes, 6100, Florianópolis, CEP 88036-700, Florianópolis, SC geinfo.sms@gmail.com

Resumo:

O absenteísmo em serviços de atendimento de saúde de média e alta complexidade do Sistema Único de Saúde está associado ao desperdício e ineficiência, uma vez que estes procedimentos envolvem grande densidade tecnológica. Nesse sentido, esse estudo busca desenvolver um modelo que possa identificar pacientes com grande probabilidade de falta em consultas especializadas, como base em aprendizagem de máquina. Para realização dessa pesquisa foram usados, como estudo-piloto, dados de atendimentos nos Centros de Especialidades Odontológicas da Secretaria Municipal de Saúde de Florianópolis entre os anos de 2016 a junho de 2018.

Palavras-chave: Absenteísmo. Consultas médicas especializadas. Faltas preditivas. Unidades de Saúde Pública de Florianópolis.

Abstract:

Absenteeism in health care services of medium and high complexity of the Unified Health System is associated with waste and inefficiency, since this procedure involves a high technological density. In this sense, this study seeks to develop a model that can identify patients with a high probability of lack in specialized consultations, as a basis in machine learning. In order to carry out this research, we used as a pilot study data, between 2016 and june 2018, from the Dental Specialties Centers of the Municipal Health Department of Florianópolis.

Keywords: Absenteeism. Specialized medical consultations. Predictive absense. Public Health Units of Florianópolis.

1 INTRODUÇÃO

O sistema único de saúde (SUS) tem convivido com problemas de financiamento desde que foi criado. Iniciativas foram tomadas para minorar, pelo menos em parte, o problema. A aprovação da CPMF[1], em 1996, e da Emenda Constitucional 29[2], em 2000, foram as mais importantes"(Piola et al., 2013) regulamentando e potencializando a atenção ao sistema único de saúde no País.

Contudo, é preciso ressaltar que os recursos destinados à saúde no Brasil, embora não ideais, não diferem de alguns países em desenvolvimento que lograram obter melhor assistência à população do que aquela que hoje, em média, é fornecida aos brasileiros (Piola et al., 2013).

Em épocas de escassez de recursos, duas ações são fundamentais para que o SUS consiga ofertar melhores serviços: 1) forçar um maior repasse para a saúde pública; e 2) melhorar os gastos em saúde, aumentando a eficiência do sistema como um todo.

Ações de advocacy para o aumento do repasse têm surgido de entidades como a ABRASCO. O aumento da eficiência, por sua vez, pode ser alcançado com o fortalecimento da atenção primária e com a redução de desperdícios na média e na alta complexidade. A atenção primária à saúde é o atendimento inicial, cujo principal objetivo é a prevenção de doenças, tratamento de agravos mais prevalentes e o direcionamento de casos menos prevalentes e mais graves para outros níveis de complexidade (SALDIVA; VERAS, 2018). Dessa forma, consegue racionalizar a utilização dos recursos em saúde, fazendo com que a entrada dos indivíduos no sistema de saúde ocorra em seu ponto mais seguro e menos dispendioso, as Unidades Básicas de Saúde (UBS). Uma boa atenção primária, deve oferecer acesso amplo e oportuno a população, fazendo com que as pessoas consigam rapidamente sanar suas principais necessidades em saúde.

Porém, a atenção primária, isoladamente, não é capaz de oferecer um cuidado integral ao cidadão, sendo, muitas vezes necessária sua articulação com serviços de saúde de média e alta complexidade. Esta articulação deve ser feita por meio de regulação, fazendo com que casos mais graves sejam atendidos de forma prioritária.

Nesse sentido, um grande problema em termos de eficiência são os desperdícios ocasionados por faltas em consultas médicas e odontológicas nos níveis mais altos de complexidade. Diferente do que ocorre na atenção primária, onde o paciente deve, preferencialmente, ser escutado no dia em que procura a UBS, reduzindo muito o número de faltas, o acesso à média e alta complexidade dá-se por encaminhamento, com marcação antecipada, fazendo com que os pacientes precisem aguardar em fila para a consulta. Desta forma, o número de faltas nestes níveis de complexidade tendem a ser maior que na atenção primária.

Em Florianópolis, por exemplo, o percentual de faltas é de 3% para consultas médicas e 12% para consultas odontológicas na atenção primária (FLORIANÓPOLIS, 2018a). Na média complexidade municipal, a mediana de faltas é de 26% (FLORIANÓPLIS, 2018b).

O absenteísmo em serviços de atendimento de média e alta complexidade vinculam-se a um alto custo, uma vez que estes são procedimento geralmente demorados e com o envolvimento de uma maior densidade tecnológica. (CONASS, 2007)

1 CPMF. Disponível em:<https://www12.senado.leg.br/noticias/entenda-o-assunto/cpmf >. Acesso em: 19 set. 2018

2 Regulamentação da Emenda Constitucional Nº 29, De 2000. Disponível em:<http://www2.camara.leg.br/orcamento-da-uniao/estudos/2012/nt14.pdf>. Acesso em: 19 set. 2018

Neste sentido, o presente trabalho visa a contribuir com a redução do desperdício no SUS, buscando produzir um modelo que possa identificar pacientes com grande probabilidade de falta às consultas especializadas, baseado em algoritmos de aprendizagem de máquina.

2 OBJETIVO

Construir um modelo de predição de possíveis faltantes à consultas especializadas no Sistema Único de Saúde.

3 PROCEDIMENTOS METODOLÓGICOS

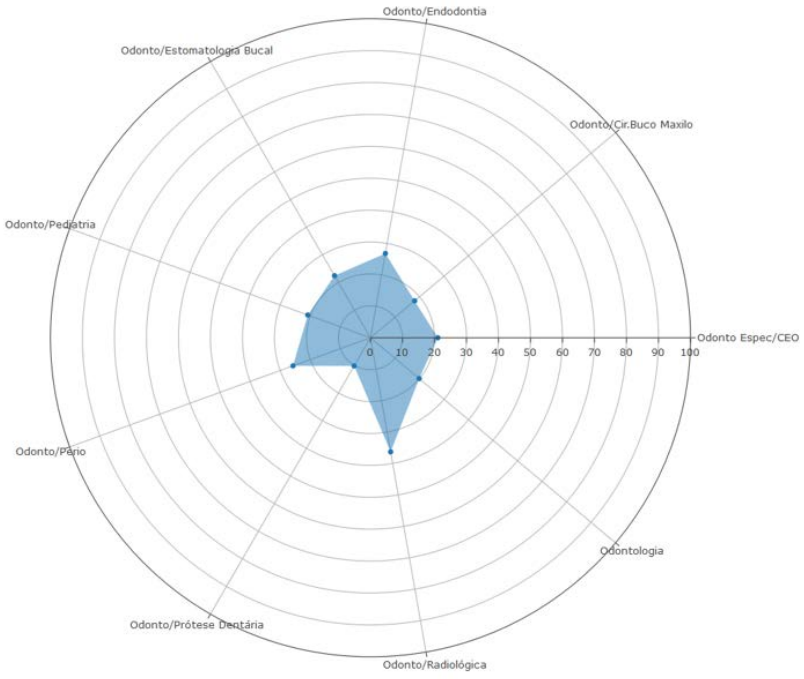
O desenvolvimento do modelo, partiu de estudo-piloto com dados secundários fornecidos pela Gerência de Inteligência e Informação da Secretaria Municipal de Saúde de Florianópolis.

Os dados correspondem a consultas agendadas no período de 1 de janeiro de 2016 a 30 de junho de 2018, realizadas nos dois Centros de Especialidades Odontológicas - CEOs do município: CEO Centro e CEO Continente. Estes serviços atendem nove especialidades odontológicas, com agendamento a partir de encaminhamentos de dentistas da atenção primária do município.

Para cada uma das especialidades, foram extraídos dos bancos: a data e o período do dia do atendimento, a Unidade de Referência (Centro de Saúde que realizou o encaminhamento), escolaridade, idade e sexo do paciente, além da indicação de primeira consulta ou retorno e de falta ou comparecimento.

Baseando-se nas especialidades com maior porcentagem de faltas neste período, foram selecionadas apenas três delas para o estudo: endodontia, periodontia e radiológica.

Gráfico 1: Percentual de faltas por especialidade nos anos 2016 /2017.



Para a etapa de predição de faltas nos atendimentos, foram aplicados os modelos de aprendizagem de máquina nos bancos de dados das especialidades. Por se tratar de um problema de classificação binária, que neste caso, será classificar se o usuário irá faltar ou não, utilizamos a matriz de confusão para aferir o preditor. Para o problema em questão, utilizamos como métrica de avaliação do preditor, a divisão do número de verdadeiros positivos pela soma dos falsos positivos e falsos negativos (VP / FP + FN). O propósito em usar essa métrica foi selecionar o algoritmo que maximiza a relação entre acertos e erros nas classificações dos faltantes.

É importante salientar, que segundo a Resolução Nº 510, de 7 de abril de 2016 (CNS, 2016), do Conselho Nacional de Saúde do Brasil, pesquisas com bancos de dados, cujas informações são agregadas, sem possibilidade de identificação individual não precisam ser submetidas a Comitês de Ética. Desta forma, não houve submissão deste projeto a um Comitê de Ética em pesquisa.

3.1 Tratamento de Dados

Os dados entre os anos de 2016 e 2017 foram usados para o treinamento dos modelos. Os modelos treinados e validados utilizando o método K-fold, com número de *folds* igual a 10. Além disso, os conjuntos de dados foram balanceados para aumentar a classe minoritária. Para os testes dos algoritmos, utilizou-se dados do ano de 2018 para verificar a capacidade de predição do melhor modelo escolhido.

3.2 Algoritmos de Aprendizado de Máquina

Utilizamos os seguintes algoritmos de aprendizagem de máquina para realização da classificação dos pacientes como faltantes ou não faltantes: Generalized Linear Model, Random Forest, k-Nearest Neighbors, Neural Network, Stacked AutoEncoder Deep Neural Network, eXtreme Gradient Boosting, Stochastic Gradient Boosting, AdaBoost Classification Trees. A implementação destes algoritmos deu-se por meio do pacote Caret do software R (R, 2018).

Algoritmos foram parametrizados da seguinte forma:

Figura 1: Parâmetros para a função de controle dos modelos.

Função "trainControl": controle que deve ser utilizado em todos os modelos para que sejam comparáveis. No código é atribuída a uma variável de nome "myControl", para poder ser utilizada da função de treino.		
Parâmetro	Valor	Descrição
method	"repeatedcv"	Dividir o banco em treinamento e teste diversas vezes. Reamostragem.
number	10	Define o número de folds (dobras) para o método de reamostragem.
summaryFunction	twoClassSummary	Calcular métricas de desempenho.
classProbs	TRUE	Computa as probabilidades dos modelos.
verboseIter	FALSE	Não imprime o log de treinamento.
savePredictions	TRUE	Quanto das predições deve ser salvo. (Tudo)
returnResamp	"all"	Quantidade das métricas que devem ser salvas. (Tudo)
sampling	"up"	Aumenta aleatoriamente a classe minoritária. Balanceamento dos dados.
allowParallel	TRUE	Permite a computação paralela

Figura 2: Parâmetros para a função de treino dos modelos.

Função "train": configuração de parâmetros para os modelos de classificação e regressão e calculo de desempenho.		
Parâmetro	Valor	Descrição
y	FALTA	Variável preditora
data	train_data	Banco no qual serão aplicados os modelos.
preProcess	c("center", "scale")	Centraliza e escala os dados de predição.
metric	ROC	Métrica para comparação dos modelos.
method	modelos[index]	Lista com índice, retorna um string com o nome do modelo.
trControl	myControl	Aplica o controle (trainControl).

Todas as análises foram executadas com o uso dos softwares estatísticos R (R, 2018) com a IDE R-Studio (RSTUDIO, 2018).

3.3 Aplicação de Ciência de Dados à previsão de Faltas

Na tentativa de encontrar padrões nos dados referentes aos anos de 2016 a junho de 2018 de atendimentos nos Centros de Especialidade Odontológicas para predizer usuários que faltam em consultas, foram feitas análises e descrições de algumas das variáveis disponíveis, sendo elas, dia da semana e período dos atendimentos, e também escolaridade e sexo dos pacientes.

4 RESULTADOS

Primeiramente, realizamos um trabalho de análise descritiva para verificar a relevância das características em cada especialidade. Posteriormente, realizamos a avaliação dos diversos modelos de aprendizado automático para verificar a eficiência na predição das faltas.

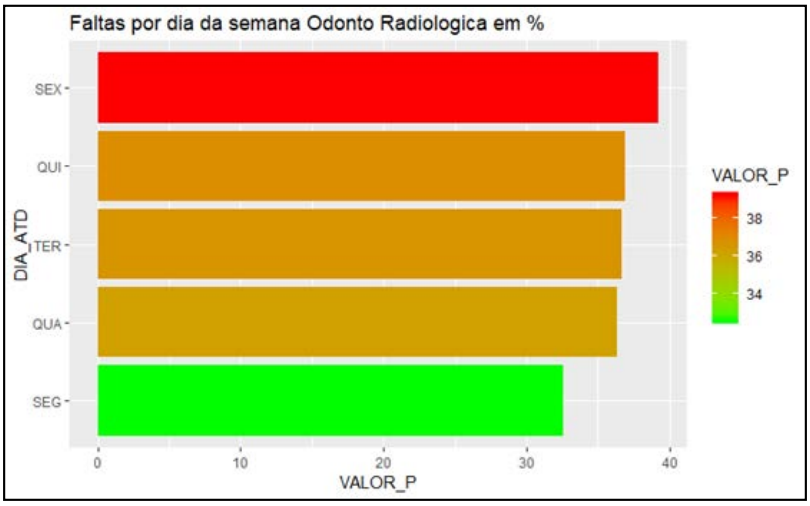
4.1 Análise Estatística Descritiva

Com as especialidades já escolhidas, obtiveram-se dados de 18.503 consultas, tendo registradas 5.944 faltas para o conjunto.

Sendo que ao separar os dados por especialidade, periodontia 2.531 atendimentos e 652 faltas, endodontia com 5.360 registros e 1.438 faltantes e por fim, radiológica com a maior quantidade de atendimentos, 10.612, assim como a maior quantidade de faltas, com 3.854.

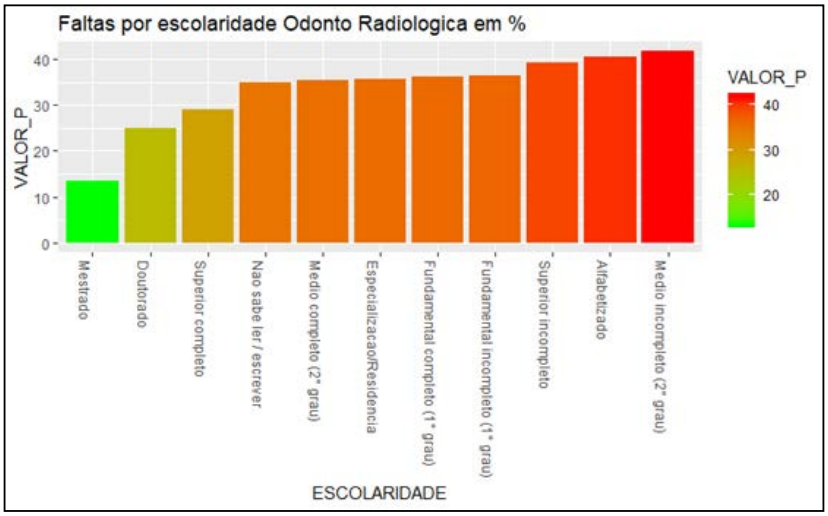
Quanto às variáveis usadas, odontologia radiológica e odontologia periodontia apresentaram sexta e sábado respectivamente como dias da semana com maior percentual de faltas, enquanto segunda feira foi o dia com menos faltantes. Na odontologia radiológica, 40% de todos os pacientes que marcaram na sexta-feira, faltaram. A endodontia foi a única das três que não apresentou sexta ou sábado como dia da semana com mais faltas, mas sim terça-feira.

Gráfico 1: Maior percentual de faltas em um dia da semana. (Radiológica)



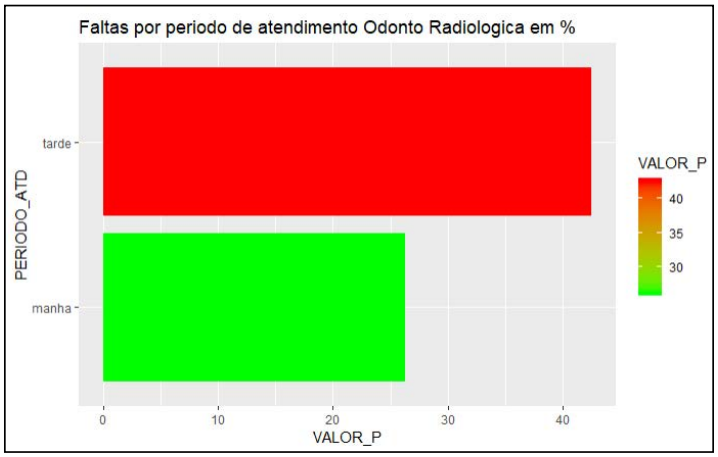
A escolaridade segue um padrão, para as três especialidades. Foi possível perceber que o de faltantes diminui com relação a pacientes que tenham ensino superior, mestrado ou doutorado. Da mesma forma, que ensino médio incompleto aparece como grau de escolaridade com o maior percentual, assim como, não alfabetizados, fundamental (completo e incompleto).

Gráfico 2: Percentual de faltas por escolaridade odontologia radiológica.



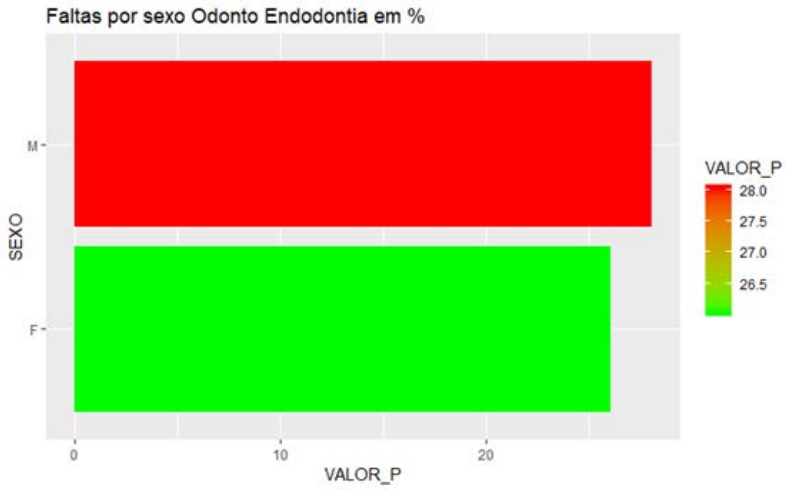
O percentual de faltas não teve muita variação na especialidade periodontia de acordo com o período de atendimento, sendo 25,8% dos pacientes que tinha consulta agendada para tarde faltará e 25,5% daqueles com consultas agendadas pela manhã. Padrão semelhante ao encontrado na Endodontia, com 28% de faltas a tarde e 25% no período matutino. A especialidade com maior discrepância de faltas por período de atendimento foi radiológica, que variou mais de 10% entre os períodos, sendo que de tarde aconteceram mais faltas.

Gráfico 3: Percentual de faltas por período em radiologia.



Não se observaram grandes variações entre o percentual de faltas entre homens e mulheres. Na Periodontia o percentual de faltas entre as mulheres tende a ser ligeiramente maior. Na Radiologia e na Endodontia homens tendem a faltar mais. Em todas as especialidades, porém, a diferença do percentual de faltas entre os sexos não passa de dois por cento.

Gráfico 4: Percentual de faltas por sexo em Endodontia



4.2 Aprendizado Automático

Em relação à seleção do modelo para predição dos faltantes, para cada uma das especialidades, usamos os dados entre 2016 a 2017 para treino e os dados de 2018 para teste. Quanto às matrizes de confusão do processo de treinamento, nenhum dos resultados apresentou especificidade boa, não passando de 0,4, o que também aconteceu para a sensibilidade, chegando à 0,739. A média entre os resultados dos três modelos com melhor desempenho de acurácia foi de aproximadamente 0,588. Entretanto, a acurácia não foi usada como método de avaliação dos algoritmos, uma vez que por se tratarem de bancos com dados não balanceados. Os modelos aplicados ao banco de dados

não balanceados predizem de forma discrepante apenas um valor da variável preditora - neste caso, que o usuário não faltará, enquanto prediz pouco que irá faltar, tornando essa métrica inválida para a avaliação. Mesmo com o balanceamento dos dados no treinamento, os resultados não melhoram na predição, a especificidade se mantém baixa e a taxa de não informação foi maior do que a acurácia para todos os resultados. Para todos os bancos, o modelo de aprendizado automático que teve o melhor desempenho foi o adaboost, alcançando sensibilidade de 0,761 e especificidade de 0,273 no banco Odonto_Endodontia (Apêndice A). Para o banco Odonto_Radiologica (Apêndice B) a sensibilidade foi a menor, chegando apenas em 0,636, mas a especificidade foi a melhor com 0,423. Já para o banco da especialidade Odonto_Perio (Apêndice C), a sensibilidade foi a maior, 0,802, mas a especificidade foi também baixa, 0,244.

5. CONCLUSÃO

Em cada cinco consultas médicas agendadas, um paciente falta e gera prejuízo de 13,4 milhões em SC em 2016 (WEISS, 2017). Apesar de implantadas medidas mais rigorosas, e exigências nas marcações de consultas ainda não impedem o absenteísmo, trazendo prejuízo à população. Nesse sentido, um modelo de previsão de faltas é imprescindível devendo diminuir os custos, pois o algoritmo pode prever se um paciente irá faltar ou não.

Este estudo-piloto indica um possível caminho para a utilização de técnicas de aprendizagem de máquina no aumento da eficiência do Sistema Único de Saúde. Os resultados preliminares dos algoritmos, podem estar relacionados à baixa capacidade preditora das *features* selecionadas. Desta forma, se faz necessário realizar mais estudos na seleção de características com maior significância no modelo de dados. Com um subconjunto de características mais significantes, é possível melhorar os preditores e por consequência, auxiliar as Secretarias de Saúde do Brasil a reduzirem o absenteísmo em consultas especializadas.

REFERÊNCIAS

ARAÚJO, Edson C.; LOBO, Maria Stella. Desafios para Sustentabilidade do Sistema Único de Saúde: **Draft for Review**. p.13, 2018

BRASIL. **Florianópolis atinge 100% de cobertura da Atenção Básica com a Estratégia de Saúde da Família.** Departamento de Atenção Primária. Ministério da Saúde. Disponível em: <http://dab.saude.gov.br/portaldab/noticias.php?conteudo=_&cod=1998>. Acessado em: 01/10/2018.

CNS. Resolução no 510, de 07 de Abril de 2016. 2016. Disponível em: <http://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf>. Acessado em: 01/10/2018.

BRASIL. (CONASS)CONSELHO NACIONAL DE SECRETÁRIOS DE SAÚDE. **Assistência de média e alta complexidade no SUS**. 2007; 248p.

FLORIANÓPOLIS. Sala de Situação da Atenção Primária. Secretaria Municipal de Saúde de Florianópolis. 2018a

_____. Sala de Situação da Regulação. Secretaria Municipal de Saúde de Florianópolis. 2018b

GONÇALVES, Cláudia Ângela et al. Estratégias para o enfrentamento do absenteísmo em consultas odontológicas nas Unidades de Saúde da Família de um município de grande porte: uma pesquisa-ação. **Ciência & Saúde Coletiva**, v. 20, p. 449-460, 2015.

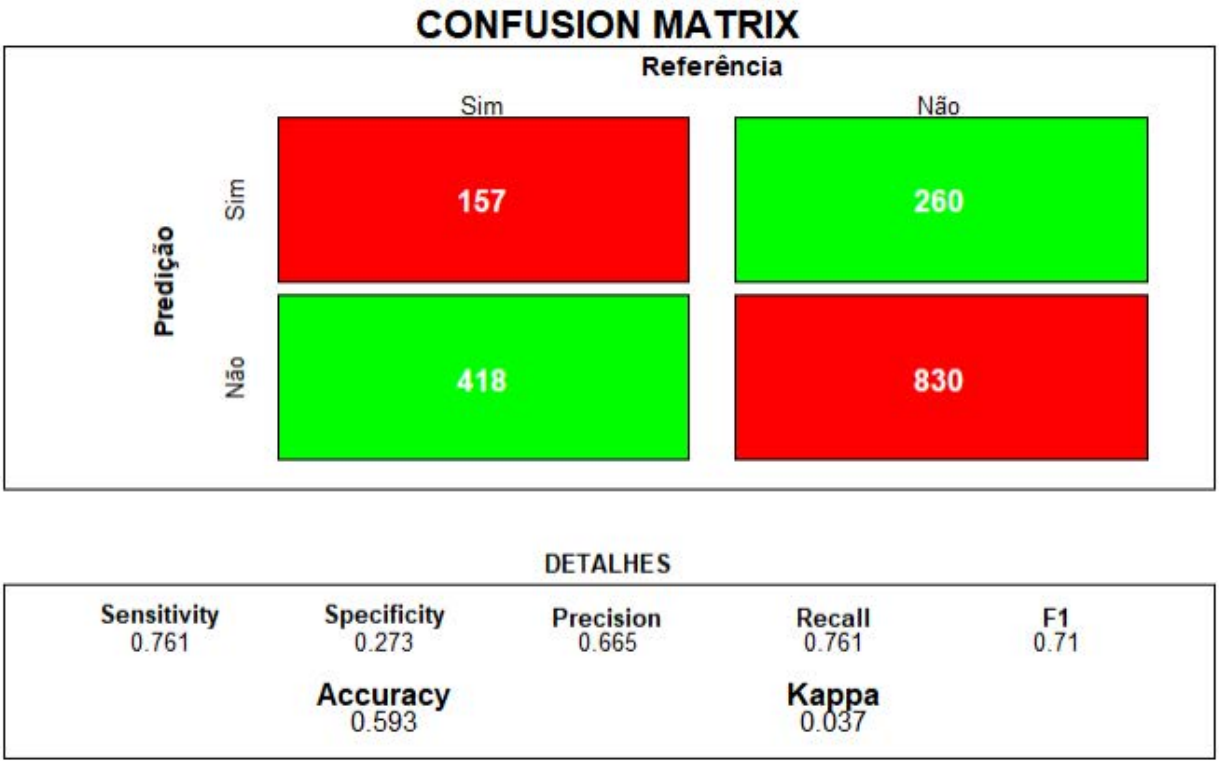
PIOLA, Sérgio Francisco et al. Estruturas de financiamento e gasto do sistema público de saúde: **A saúde no Brasil em 2030-prospecção estratégica do sistema de saúde brasileiro: estrutura do financiamento e do gasto setorial** [online]. Rio de Janeiro: Fundação Oswaldo Cruz, v. 4, p. 19-70, 2013.

R – **A language and environment for statistical computing R Foundation Statistical Computing**, 2018. Disponível em: <https://www.r□project.org>. Acessado em: 01/10/2018.

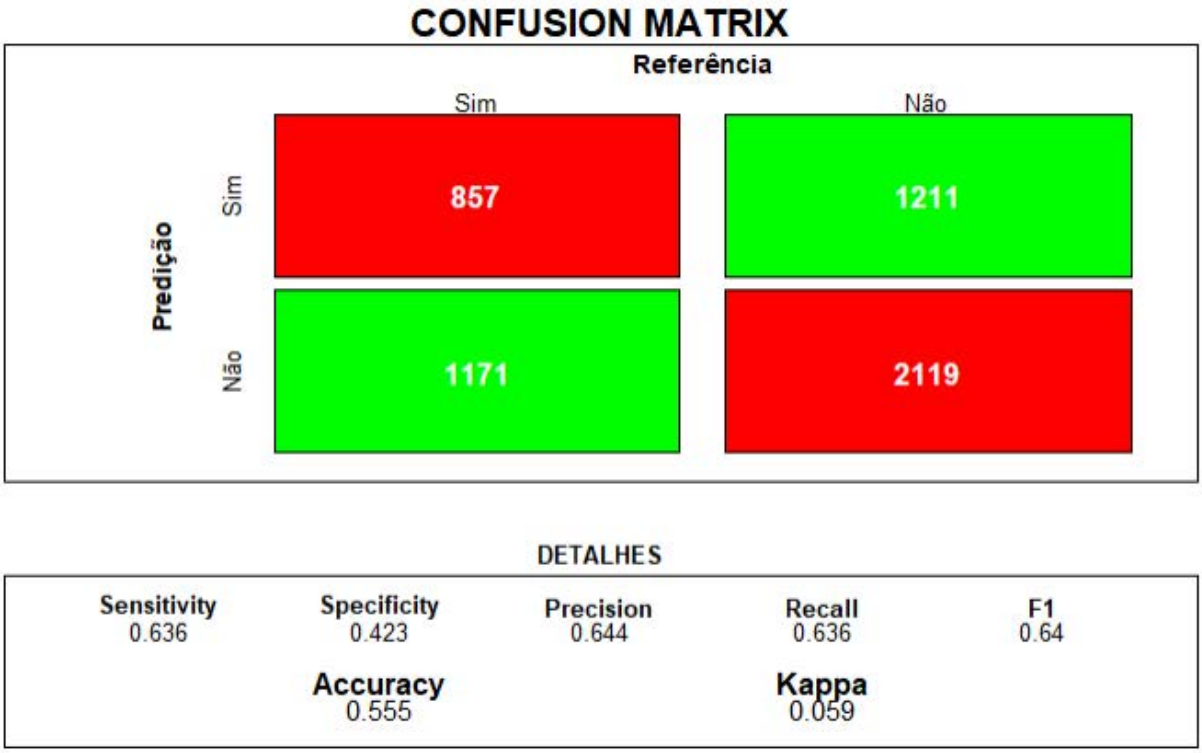
RStudio: **integrated development for R**
RStudio, Inc, 2018. Disponível em: <http://www.rstudio.com/>. Acessado em: 01/10/2018.

SALDIVA, Paulo Hilário Nascimento; VERAS, Mariana. Gastos públicos com saúde: breve histórico, situação atual e perspectivas futuras. **Estudos Avançados**, v. 32, n. 92, p. 47-61, 2018.

Apêndice A – Resultados banco OdontoEndodontia

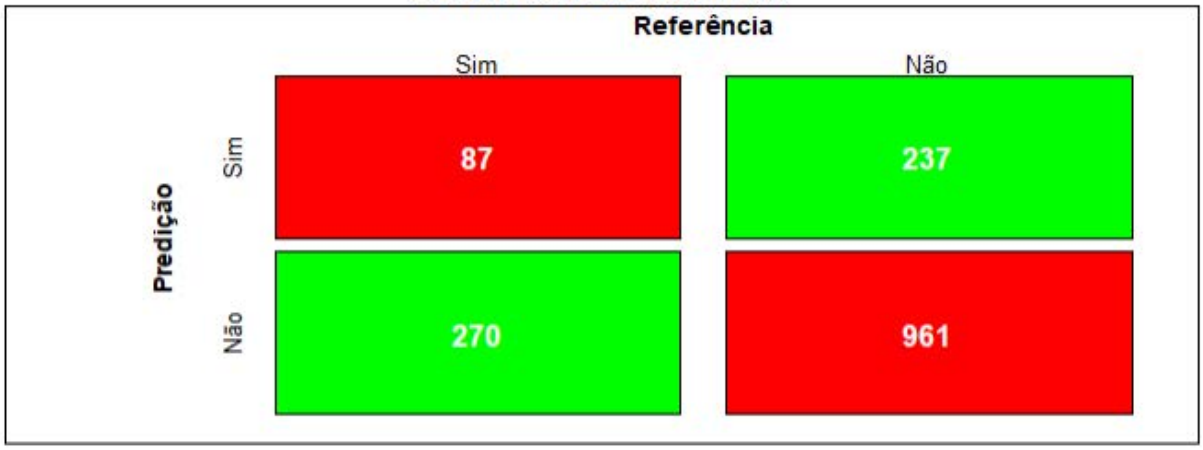


Apêndice B - Resultados banco OdontoRadiologica



Apêndice C - Resultados banco OdontoPerio

CONFUSION MATRIX



DETALHES

Sensitivity	Specificity	Precision	Recall	F1
0.802	0.244	0.781	0.802	0.791

Accuracy	Kappa
0.674	0.047

PROTEÇÃO AUTORAL DE DADOS:
Uso de Blockchain

AUTHORAL PROTECTION OF DATA:
Use of Blockchain

Elizabete Cristina de Souza Aguiar Monteiro[1], Jacquelin Teresa Camperos Reyes [2], Elaine Parra Affonso[3], Ricardo César Gonçalves Sant'Ana[4]
(1) Universidade Estadual Paulista, Marília, ecsamonteiro@gmail.com
(2) Universidade Estadual Paulista, Marília, jcamperosreyes@gmail.com
(3) FATEC, Presidente Prudente, elainepff@gmail.com
(4) Universidade Estadual Paulista, Marília, ricardo.santana@unesp.br

Resumo:
A tecnologia *blockchain* pode apresentar potencial para proteger o direito autoral de dados quando estes estão envolvidos em ambientes digitais. Com o intuito de explicitar como *blockchain* tem sido abordada para questões de proteção do direito autoral de dados, foi realizada uma revisão sistemática de literatura, junto com análise de conteúdo, em trabalhos publicados na base de dados *Scopus*. Os resultados demonstram que o uso da tecnologia *blockchain* para a proteção de direitos autorais é pouco recorrente nos 80 estudos analisados, apenas oito apresentam aplicação dessa tecnologia para proteção autoral de dados. Entre as aplicações que emergem nesse cenário há plataformas como Mycelia, criação de marca d´água e contratos inteligentes. Conclui-se que o uso de blockchain no contexto de proteção de dados autorais não é foco das pesquisas recuperadas, situação que torna essa temática uma oportunidade para pesquisas em diversas áreas do conhecimento.
Palavras-chave: Direitos autorais. Blockchain. Propriedade intelectual. Dados.

Abstract: Blockchain technology may have the potential to protect copyright when data are involved in digital environments. In order to explain how blockchain has been approached for copyright protection issues, a systematic review of the literature, together with content analysis, was carried out in published works in the Scopus database. The results show that the use of blockchain technology for copyright protection is not very recurrent in the 80 studies analyzed, only eight present application of this technology for data protection. Among the applications that emerge in this scenario are platforms like Mycelia, watermark creation and smart contracts. It was observed that blockchain is involved with studies and applications in several areas of knowledge, marking this topic as a fertile field for research in the scope of protection and verification of data copyright.
Keywords: Copyright. Blockchain. Intellectual property. Data.

1 INTRODUÇÃO

O avanço na ciência e na produção de dados em diversas atividades da sociedade podem proporcionar oportunidades no âmbito da colaboração internacional, geração de conhecimentos e progresso da ciência, porém, apresenta desafios quando relacionados à reutilização e a proteção autoral dos dados.

O compartilhamento ou publicação de dados oferecem vantagens para os cientistas pois, quando os pesquisadores publicam seus dados, o seu trabalho pode ser compartilhado entre as fronteiras regionais, nacionais e internacionais (TENOPIR et al., 2015).

Para o pesquisador que busca melhorar seu *status* profissional e seu reconhecimento perante seus pares (agradecimentos e citações), o compartilhamento de seus dados propicia oportunidades para ser reconhecido como o autor, ser citado por membros mais proeminentes de seu campo de atuação e receber o reconhecimento formal (KIM; STANTON, 2013, 2016). Da mesma forma, a divulgação de dados fora do contexto científico impacta positivamente nas carreiras dos profissionais.

Os fatores apresentados instigam o compartilhamento dos dados, no entanto, os custos na geração ou coleta, no tratamento e na gestão de dados são relativamente dispendiosos e os detentores dos dados têm recorrido a estratégias para proteger a autoria dos dados. Um recurso utilizado para questões de proveniência e garantia da autoridade de dados é a tecnologia *blockchain*.

Segundo Kim e Laskowski (2018) *blockchain* é um registro público baseado em uma arquitetura distribuída, uniformemente replicado em todos os seus nós, à prova de adulteração, que fornece um método para estabelecer a existência de uma transação em um momento específico, podendo ser verificada por qualquer parte interessada, incluindo o rastreamento da proveniência dos dados.

Tecnicamente, *blockchain* é um sistema distribuído público que contém todas as transações que já foram executadas no sistema. No contexto da gestão de dados de pesquisa, *blockchain* permite gerir esses dados de forma descentralizada, com a possibilidade da auto-regulação por meio de uma infraestrutura compartilhada para transações ou fluxos dos dados. A tecnologia também permite que pesquisadores interajam com o conteúdo em qualquer etapa do ciclo de vida dos dados. Ainda, interessante ressaltar que, o uso de *blockchain* nas questões vinculadas à proteção do direito autoral:

> [...] permite, por um lado, a criação de um registro permanente dos acordos entre os autores dos dados e os reutilizadores, com a possibilidade de verificar a conformidade a qualquer momento e, por outro lado, um maior nível de granularidade na definição das condições de reutilização. (PANESCU; MANTA, 2018)

Assim, o registro é fator significativo, na medida em que confere a autenticidade do protocolo original e a segurança jurídica ao autor, facilita a comprovação da criação da obra, assegura o direito autoral na publicação e a obtenção de um certificado de registro. Ainda, para Karafiloski (2017) quando se tem o registro de data e hora das transações, é possível obter prova no tribunal, no caso de uma disputa de propriedade.

2 OBJETIVOS

No contexto desta pesquisa busca-se investigar a aplicação da tecnologia *blockchain* para proteção do direito autoral de dados e relatar como está sendo utilizada.

3 PROCEDIMENTOS METODOLÓGICOS

A metodologia deste trabalho consiste em uma revisão sistemática de literatura a fim de verificar como a tecnologia *blockchain* tem sido aplicada para comprovação de direitos autorais em dados. As estratégias utilizadas nesta metodologia foram baseadas no protocolo de revisão sistemática de Kitchenham (2004). Para tanto, realizaram-se as seguintes estratégias:

a) Base de dados: Foi escolhida a *Scopus* por fornecer um serviço eficaz na obtenção de informações de produção científica (NOBLEJAS; RODRIGUEZ, 2014), envolvida com diversas áreas de conhecimento, visto que a temática deste trabalho abarca questões da Computação, do Direito e da Ciência da Informação;
b) Termo descritor: A busca foi realizada no dia 2 de outubro de 2018, por meio da seguinte estratégia de busca: (TITLE-ABS-KEY (BLOCKCHAIN) AND (TITLE-ABS-KEY ("intellectual property" OR copyright) AND TITLE-ABS-KEY (blockchain)). Os termos foram escolhidos considerando o foco da pesquisa na tecnologia *blockchain* como opção para proteção do direito autoral de dados;
c) Tipos de documentos: Foram analisados todos os documentos recuperados compostos por artigos de periódicos, capítulos de livros e artigos publicados em anais de eventos;
d) Ano de publicação: Foram considerados para a análise todos os documentos recuperados mediante a expressão de busca utilizada, assim, a coleta não foi delimitada por valor temporal;
e) Critério de exclusão: Não foi excluído nenhum documento recuperado;
f) Processo de seleção: Após a expressão de busca ser submetida à base de dados *Scopus,* os atributos referentes a cada documento recuperados foram armazenados em *software* de planilhas, extraindo as seguintes informações: Ano de publicação, título do artigo, quantidade de citações, resumo e palavras-chaves;
g) Processo de análise: Foi aplicada Análise de conteúdo, adotando a análise léxica, tendo como unidade de registro a palavra, e utilizando a regra de enumeração "presença" (BARDIN, 1977). Para tanto, foi utilizado a ferramenta *textalyser*, que possibilita realizar a análise estatística de um texto, permitindo descobrir palavras-chave. Posteriormente utilizou-se a linguagem R para geração da nuvem de palavras;
h) Mediante a leitura dos resumos realizou-se a análise categorial proposto no método Análise de conteúdo para verificar a principal abordagem no uso de *blockchain* no contexto de proteção de dados autorais. Esta atividade permitiu a construção do quadro no Apêndice A com as seguintes categorias: tipo de dado, técnica/processo e descrição do processo.

Foram utilizados quadros e figuras para representar os dados extraídos dos documentos, com a finalidade de demonstrar a frequência das palavras nos textos e a quantidade de trabalhos por ano. Por meio desses dados é possível inferir sobre o objeto de estudo deste trabalho.

4 RESULTADOS

Foram recuperados 80 trabalhos na base de dados *Scopus* de acordo com a estratégia de busca *blockchain* and *"copyright or intellectual property" and data*. Após a coleta, os resumos foram agrupados em um único arquivo, resultando em um arquivo com 14 páginas de texto. Este arquivo foi inserido na ferramenta *textalyser* para realizar a análise de contagem de palavras e frequência.

O Quadro 1 apresenta as palavras com maior número de ocorrências, resultado da análise léxica realizada pela ferramenta *textalyzer* nos resumos e nas suas palavras-chaves. Observa-se que o termo descritor "dados" e "*blockchain"* são os mais recorrentes nos trabalhos recuperados, incluindo os termos tecnologia, sistema, *bitcoin* e segurança. Interessante ressaltar que termos relacionados à proteção dos direitos autorais não emergem entre os 20 primeiros termos recorrentes.

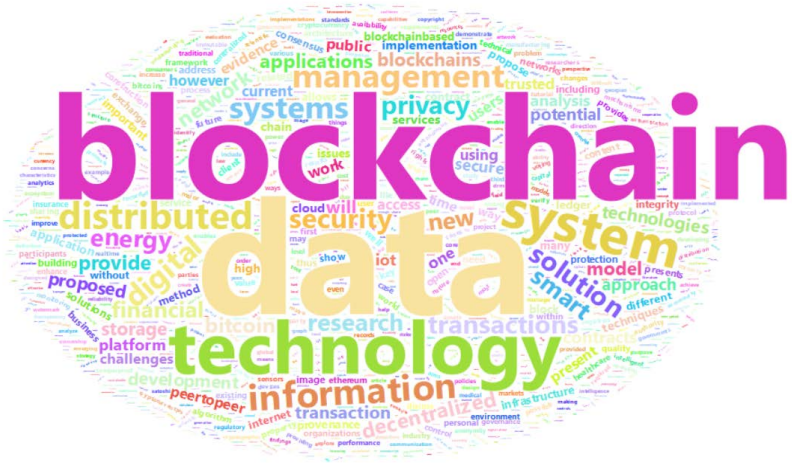
Quadro 1: Ocorrências de palavras nos resumos e nas palavras-chaves dos documentos

Resumo			**palavras-chaves**		
Palavra	Qtd	Freq	Palavra	Qtd	Freq
data	164	2.7%	data	79	4.5%
blockchain	160	2.6%	systems	52	3%
technology	77	1.3%	block	35	2%
system	56	0.9%	information	34	1.9%
distributed	39	0.6%	chain	34	1.9%
information	37	0.6%	management	33	1.9%
management	34	0.6%	distributed	31	1.8%
peer	34	0.6%	blockchain	29	1.7%
systems	32	0.5%	peer	27	1.5%
digital	31	0.5%	electronic	26	1.5%
bitcoin	30	0.5%	privacy	26	1.5%
solution	29	0.5%	security	25	1.4%
security	28	0.5%	digital	25	1.4%
privacy	28	0.5%	money	22	1.3%
transactions	26	0.4%	internet	19	1.1%
energy	26	0.4%	storage	19	1.1%
smart	25	0.4%	things	15	0.9%
research	24	0.4%	computing	14	0.8%
network	23	0.4%	computer	14	0.8%
financial	23	0.4%	networks	14	0.8%

Fonte: Dados da pesquisa.

Ao realizar mineração de texto na linguagem R para proporcionar a visualização dos dados mais frequentes nos resumos dos trabalhos recuperados, obteve-se a Figura 1, que também corrobora com os dados demonstrados na análise anterior (QUADRO 1).

Figura 1: Termos com maior ocorrência

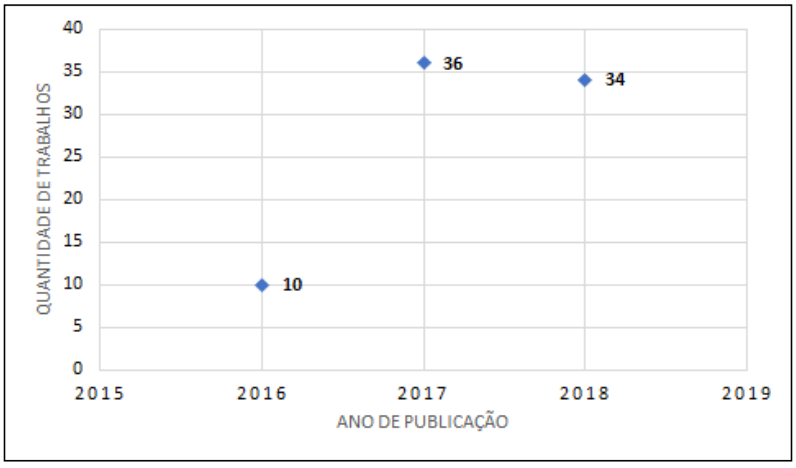


Fonte: Dados da pesquisa.

Com as principais ocorrências de termos nas palavras-chaves dos documentos (Quadro 1) pode-se inferir que prevalece questões voltadas a segurança dos dados, efetivado pelos termos "security", "privacy", ressalta-se também que o uso de *blockchain* pode estar voltado para proteção de transações que envolvem dinheiro digital, tal como *bitcoin*. Assim, nos 20 termos com mais ocorrências nas palavras-chaves não aparece termos relacionados à proteção de direitos autorais.

Observa-se na figura 2 que os documentos recuperados relacionados à proteção autoral de dados e *blockchain* são publicados na base de dados *Scopus* a partir do ano de 2016, isso demonstra que a temática é recente na literatura e apresenta um crescimento de 260% do ano 2016 para o ano 2017.

Figura 2: Ocorrência de publicação/ano



Fonte: Dados da pesquisa.

No apêndice A estão relacionados os trabalhos recuperados que relatam sobre o uso do *blockchain* para proteção de direitos autorais. Na primeira e segunda coluna estão os nomes dos autores e os títulos dos documentos respectivamente, na terceira coluna estão indicados os tipos

de dados abordados nos documentos, na quarta e quinta coluna estão as técnicas ou os processos aplicados com a tecnologia *blockchain* e, o resumo dos processos apresentados pelos autores.

Observa-se que *blockchain* é aplicada em dados identificados pelos autores como: dados de pesquisa, dados digitais ou simplesmente dados, explicitando o contexto dos quais os dados estão inseridos (MA; HUANG; GAO, 2018; PĂNESCU; MANTA, 2018), músicas e vídeos (GAO; NOBUHARA, 2017; O'DAIR; BEAVEN, 2017), imagens de obras de arte (ZHAOFENG; WEIHUA; HONGMIN, 2018) e *timestamp* (GIPP et al.2017).

Os artigos descrevem e discutem técnicas e/ou processos utilizados com o *Blockchain* como a plataforma MyCelia para música, o CryptSubmite para manuscritos, contratos inteligentes e marcas d'água para obras de arte (APÊNDICE A).

Dentre os principais benefícios e oportunidades relatados destacam-se os relacionados à precisão e acessibilidade aos dados de direitos autorais, transparência nos processos, meios para evitar a falsificação de dados e o gerenciamento de direitos autorais.

5 CONSIDERAÇÕES FINAIS

A revisão sistemática de literatura na base de dados Scopus revelou que apesar de serem recuperados 80 trabalhos, apenas oito apresentam a aplicação de *blockchain* para proteção autoral de dados. Mesmo os trabalhos que discutem aspectos relacionados a proveniência de dados, esses descrevem a facilidade da integração de dados de diversas fontes, a entrega de informações pelas fontes, a origem dos dados, a verificação de adulteração dos dados identificando em quais dados foram feitas as transformações, no entanto, não vinculam com o direito autoral.

Conclui-se que o uso de blockchain no contexto de proteção de dados autorais não é foco da maioria das pesquisas recuperadas, situação que torna essa temática uma oportunidade para pesquisas em diversas áreas do conhecimento, visto que, ao utilizá-la pode-se contribuir na proteção e comprovação do direito autoral de dados.

REFERÊNCIAS

BARDIN, L. **Análise de conteúdo**. 4. ed. Lisboa: Edições 70, 1977.

CHEN, J.; XUE, Yi. Bootstrapping a blockchain based ecosystem for big data exchange. In: Big Data (BigData Congress), 2017 IEEE International Congress on. IEEE, 2017. p. 460-463.

GAO, Y.; NOBUHARA, H. A Decentralized Trusted Timestamping Based on Blockchains. **IEEJ Journal of Industry Applications**, Tokyo, v. 6, n. 4, p. 252-257, 2017.

GIPP, B. et al. CryptSubmit: introducing securely timestamped manuscript submission and peer review feedback using the Blockchain. In: Proceedings of the ACM/IEEE joint conference on digital libraries, 2017, Toronto. **Procedings…** 2017. Disponível em: <https://ieeexplore.ieee.org/document/7991588 >. Acesso em: 20 set. 2018.

KARAFILOSKI, E.; MISHEV, A. Blockchain solutions for big data challenges: A literature review. In: Smart Technologies, IEEE EUROCON International Conference on IEEE, 17., 2017. **Proceedings…**2017. p. 763-768. Acesso em: 20 set. 2018.

KIM, H. M.; LASKOWSKI, M.Toward an ontology‐driven blockchain design for supply‐chain provenance. **Intelligent System in Accounting Finance e Management,** Hoboken, v. 25, n. 2, p. 18-27, Jan./Mar. 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/isaf.1424>. Acesso em: 20 set. 2018.

KIM, Y.; STANTON, J. M. Institutional and individual influences on scientists' data sharing practices. **Journal of Computational Science Education**, v. 3, p. 47-56, 2013. Disponível em: <https://www.researchgate.net/publication/239522383_Institutional_and_Individual_Influences_on_Scientists'_Data_Sharing_Practices>. Acesso em: 29 set. 2018.

KIM, Y.; STANTON, J. M. Institutional and individual factors affecting scientists' data‐sharing behaviors: A multilevel analysis. **Journal Of the Association for Information Science and Technology**, v. 67, n. 4, p. 776–799, 2016. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/asi.23424>. Acesso em: 29 set. 2018.

KITCHENHAM, B. **Procedures for performing systematic reviews**: joint technical report. Keele: Keele University, 2004.

MA, Z.; HUANG, W.; GAO, H. Secure DRM Scheme Based on Blockchain with High Credibility. Chinese Journal of Electronics, v. 27, n. 5, p. 1025-1036, 2018.

NOBLEJAS, C. J.; RODRÍGUEZ, A. P. Recuperación y visualización de información en Web of Science y Scopus: una aproximación práctica. **Investigación Bibliotecológica:** archivonomía, bibliotecología e información, México, v. 28, n. 64, p. 15-31, 2014. Acesso em: 20 set. 2018.

O'DAIR M., BEAVEN Z. The networked record industry: How blockchain technology could transform the record industry. **Strategic Change**, Hoboken, v. 26, n. 5, p. 471-480, Sept. 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/jsc.2147>. Acesso em: 20 set. 2018.

PÃNESCU, A-T.; MANTA, V. Smart Contracts for Research Data Rights Management over the Ethereum Blockchain Network. **Science & Technology Libraries**, p. 1-11, 2018.

TENOPIR C. et al. (2015) Changes in data sharing and data reuse practices and perceptions among scientists worldwide. **PLoS ONE**, San Francisco, v. 10, n. 8, Aug. 2015. Disponível em: <https://doi.org/10.1371/journal.pone.0134826>. Acesso em: 16 fev. 2018.

ZHAOFENG, M.; WEIHUA, H.; HONGMIN, G. A new blockchain-based trusted DRM scheme for built-in content protection. **EURASIP:** Journal on Image and Video Processing, v. 2018, n. 91, p. 1-12, Dec. 2018. Disponível em: <https://link.springer.com/article/10.1186/s13640-018-0327-1>. Acesso em: 20 set. 2018.

Apêndice A – Uso de blockchain para proteção autoral de dados

AUTORES	TÍTULO DO ARTIGO	TIPO DE DADO	TÉCNICA/PROCESSO	DESCRIÇÃO DO PROCESSO
O'DAIR; BEAVEN	The networked record industry: How blockchain technology could transform the record industry	Música	Plataforma Mycelia; banco de dados universal e único de direitos autorais de música; Aurovine; Global Repertoire Database (GRD); contratos inteligentes; micropagamentos; proof-of-work	Precisão e acessibilidade dos dados de direitos autorais; velocidade de pagamentos de royalties; e transparência da cadeia de valor.
GIPP et al.	CryptSubmit: Introducing Securely Timestamped Manuscript Submission and Peer Review Feedback Using the Blockchain	Timestamp	CryptSubmit - sistema de submissão de manuscritos para a publicação acadêmica	Sistema que cria automaticamente um timestamp descentralizado, à prova de falsificação e publicamente verificável para cada manuscrito submetido
PANESCU; MANTA	Smart Contracts for Research Data Rights Management over the Ethereum Blockchain Network	Dados de pesquisa	Solidity Contrato inteligente	Permite aos autores supervisionar como está sendo acessado e reutilizado.
MA; HUANG; GAO	Secure DRM scheme based on blockchain with high credibility	Dados digitais	Infraestrutura BaaS com plataforma DRM	Criação de blocos gênesis, definição de mecanismos de consenso, inicialização e execução de nós, gerenciamento de wallet, gerenciamento de endereços, blockchain explorer
CHEN; XUE	Bootstrapping a Blockchain Based Ecosystem for Big Data Exchange	Dados	Ecossistema para troca de dados	Visa cultivar um ecossistema, dentro do qual todos os participantes podem cooperar para trocar dados de maneira peer-to-peer
GAO; NOBUHARA	A decentralized trusted timestamping based on blockchains	Dados digitais (documentos, música e vídeos	Registro de data e hora	Usando blockchain expandindo o espaço de armazenamento de 40 bytes para N * 20 bytes.
ZHAOFENG; WEIHUA; HONGMIN	A new blockchain-based trusted DRM scheme for built-in content protection	Imagem de obras de arte	Marca d'água com esquema DRMChain	Esquema de gerenciamento de direitos digitais de imagens de obras de arte para detecção de uso incorreto, incorporamos informações corretas de obras de arte, como autor, RightHolder, data e local

Fonte: Dados da pesquisa.

RECONHECIMENTO DE ENTIDADES NOMEADAS EM RELATÓRIOS DE INTELIGÊNCIA FINANCEIRA

ENTITY NAMED RECOGNITION IN FINANCIAL INTELLIGENCE REPORT

Jairo Santana[1], Diefferson K. Môro[1], Rogério de Aquino Silva[1], Vinicius F. C. Ramos[1], Gustavo Medeiros de Araújo[1]
(1) Universidade Federal de Santa Catarina, R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900
{jairo.santana, differson.moro, rogerriomp}@gmail.com, {v.ramos, gustavo.araujo}@ufsc.br

Resumo:
O reconhecimento de entidades nomeadas é uma das subáreas do processamento de linguagem natural, mineração de textos e aprendizado de máquinas. Todas essas áreas fazem parte da grande área da inteligência artificial, muito utilizada em diversos problemas práticos do nosso dia a dia. Uma das competências da Polícia Federal é a investigação de crimes financeiros, em especial, a lavagem de dinheiro. Dentre os problemas encontrados na investigação policial, destacamos a análise dos Relatórios de Inteligência Financeira (RIF), escritos em português do Brasil, que são gerados pelo Conselho de Controle de Atividades Financeiras. O objetivo desta análise é identificar os atores envolvidos em esquemas de lavagem de dinheiro, mas, dependendo da complexidade do esquema, a identificação, por exemplo, desses atores e suas relações (sociedades, parentescos, "laranjas", empresas "fantasmas", etc) em um relatório, pode demandar um tempo significativo do policial envolvido na investigação. Este trabalho, portanto, visa apresentar resultados iniciais da automatização do reconhecimento de entidades nomeadas, importantes para a investigação policial, em RIFs. Identificamos, na literatura, uma grande lacuna para esse tipo de solução em textos em português. Os nossos resultados, ainda preliminares, demonstram que as ferramentas e os dados utilizados para o treinamento ainda precisam ser melhor trabalhados para que estes sejam mais significativos. Pudemos perceber que com poucos dados de treinamento conseguimos aumentar a precisão do reconhecimento de entidades de 14 para 27% e, em um teste com o framework RASA NLU, aumentamos a precisão para 60,98% de entidades reconhecidas corretamente, muito aquém dos 90% encontrados na literatura para outros idiomas.
Palavras-chave: Reconhecimento de Entidades Nomeadas. Mineração de Texto. Relatório de Inteligência Financeira. Processamento de Linguagem Natural.

Abstract:
The named entity recognition is a subarea of natural language processing, text mining, and machine learning. These areas are part of the artificial intelligence area, very used in different kind of daily practical problems. One of the competencies of the Brazilian Federal Police is to investigate financial crimes, especially money laundering. Among the problems encountered in the police investigation, we highlight the analysis of the Financial Intelligence Reports, written in Brazilian Portuguese, which are generated by the Financial Activities Control Council. The aim of this analysis is to identify the actors involved in money laundering schemes, but depending on the complexity of the scheme, the identification, for example, of these actors and their relationships (societies, kinship, "oranges", "ghost" companies, etc.) in a report, may require significant time from the police man involved in the investigation. The main objective of this paper is to present initial results of the automation of named entity recognition, important for the police investigation, in Financial Intelligence Reports. We identified, in the literature, a large gap for this type of solution in Portuguese texts. Our preliminary results demonstrate that the tools and data used for training still need to be better explored to make them more meaningful. We could see that, with a few training datasets, we were able to increase the accuracy of the recognition of entities from 14 to 27% and, using the Rasa NLU framework, we got a 60.98% precision, very below the 90% found in the literature for other languages.
Keywords: Named Entity Recognition. Text Mining. Financial Intelligence Report. Natural Processing Language.

1 INTRODUÇÃO

Entre as diversas atribuições da Polícia Federal, encontra-se a investigação de crimes de lavagem de dinheiro (BRASIL, 1998), que pode ser definido como "a atividade de investir, ocultar, substituir ou transformar e restituir o dinheiro de origem ilícita aos circuitos econômico-financeiros legais, incorporando-o a qualquer tipo de negócio como se fosse obtido de forma lícita" (CALLEGARI, 2001 apud LUSTOSA, 2009). Lustosa (2009) refere-se a esse crime como uma maneira genérica de ocultação da origem de dinheiro ou bens advindos de atividades delitivas, seja por um processo ou um conjunto de operações, e a sua respectiva integração ao sistema financeiro por meio de operações capazes de converter o dinheiro sujo em dinheiro limpo.

Esse complexo esquema de lavagem de dinheiro pode ser dividido em três fases: ocultação, dissimulação e integração. A ocultação consiste basicamente em esconder e afastar o ativo da origem ilícita, para que se possa evitar que seja rastreado. Uma das maneiras de fazê-la é realizar diversas transações com valores fracionados, que desobriga a sua comunicação às autoridades financeiras (MENDRONI, 2015). A próxima etapa é a dissimulação, que tem por fim disfarçar a origem criminosa dos valores, camuflando evidências através de uma série de complexas transações financeiras internacionais em países que não cooperam com o combate à lavagem de dinheiro, os chamados *paraísos fiscais* (BRAGA, 2010). Por fim, a fase de integração é onde se tem os benefícios dos ativos como se fossem lícitos, seja através da compra de bens ou no investimento em empresas comerciais criadas e operando de forma legal.

O Conselho de Controle de Atividade Financeiras (COAF) é o responsável pela análise das evidências de crime de lavagem de dinheiro ao ser comunicado de operações suspeitas. Confirmadas as evidências, haverá uma troca de informações com as autoridades competentes (MARQUES, 2014), da forma que está descrito no artigo 15 da Lei nº 9.613/1998.

A Polícia Federal, o Ministério Público Federal, as polícias civis e os ministérios públicos estaduais são consideradas as autoridades competentes para receberem os informes do COAF, conforme o contexto da operação. Esses órgãos poderão bloquear a operação financeira suspeita, iniciar uma investigação criminal e, até mesmo, propor a ação penal (ARAS, 2007). O meio pelo qual o COAF disponibilizará seus informes às autoridades competentes é através do Relatório de Inteligência Financeira (RIF).

Os RIFs enviados para Polícia Federal são recebidos pela área de repressão a crimes financeiro e inicialmente avaliados, detectando qual será a delegacia responsável pela análise do relatório. O RIF será analisado detalhadamente de forma a identificar se realmente possuem indícios suficientes que retratam uma ação criminosa e, por fim, a instauração de um procedimento investigativo formal (POLÍCIA FEDERAL, 2013). O RIF é um documento escrito em linguagem natural, sem um padrão de estrutura e apresenta um relato conforme a compreensão do analista do COAF no formato digital PDF – *Portable Document Format*.

Atualmente, as áreas de inteligência da PF analisam o RIF de modo manual, identificando entidades, valores, operações realizadas e possíveis vínculos entre as entidades (POLÍCIA FEDERAL, 2013). Conforme a complexidade do relato no RIF, apenas uma leitura é feita para se realizar a análise ou, nos casos mais complexos, suas informações deverão ser tabuladas, armazenadas em planilhas eletrônicas ou em banco de dados, possibilitando sua leitura por softwares analíticos que permitem

gerar diagramas de relacionamentos entre as entidades, para, finalmente, facilitar a detecção de Organizações Criminosas. É primordial que a detecção e tabulação de entidades e outras informações aconteça com a máxima acurácia possível para não haver a possibilidade de se comprometer uma correta análise na fase posterior.

Nesse contexto, esse estudo tem como objetivo automatizar o processo de reconhecimento de entidade nomeada (REN), em especial, a identificação dos atores envolvidos nos RIFs utilizando ferramentas e técnicas de Processamento de Linguagem Natural (PLN), mineração de textos e aprendizado de máquina.

2 TRABALHOS RELACIONADOS

Pires (2017) apresenta em sua dissertação de mestrado alguns resultados sobre abordagens e configurações de ferramentas e corpus anotados em língua portuguesa para a pesquisa em um sistema de busca de notícias, da Universidade do Porto. O sistema de notícias chama-se SIGARRA. Segundo o autor, a ferramenta CoreNLP utilizada com o corpus HAREM obteve o melhor resultado para o Reconhecimento de Entidades Nomeadas (REN), chegando a 56% de precisão. Após essa identificação, o autor criou um conjunto de testes com 905 notícias anotadas e 12644 entidades anotadas. Nesse caso, o sistema implementado atingiu 86,86% em relação à *f-measure*.

Fonseca et al (2017) apresentam uma ferramenta para o REN baseada no *NameFinder*, uma das classes da ferramenta OpenNLP. Os autores utilizam o Corpus Amazônia para treinar o modelo e o Corpus HAREM para validar o modelo. Os resultados apontam para uma precisão de 58,65%, com recall de 56,60% e f-measure de 57,61% para as entidades do tipo Pessoa. Os autores ainda mostram resultados de diversas outras ferramentas de NER para o português e concluem que esses resultados são compatíveis com essas ferramentas, mas ainda aquém dos resultados encontrados em outros idiomas.

O diferencial do nosso trabalho é o uso de um contexto determinado, reconhecimento de entidades nomeadas em relatórios de inteligência financeira, o que poderá nos auxiliar na identificação dessas entidades. Além disso, propomos a comparação de outros modelos para o treinamento que não os apontados na literatura.

3 PROCEDIMENTOS METODOLÓGICOS

O reconhecimento de entidades nomeadas (REN) é um dos principais elementos do PLN. Ele é essencial para várias etapas do PLN, dentre elas a classificação de uma frase ou a checagem de vínculos entre as entidades.

Neste contexto, este trabalho fez uma análise de algumas ferramentas e técnicas utilizadas para a REN em textos de inteligência financeira (RIF) escritos em português e compartilhados pelo COAF.

A primeira etapa do nosso processo é a identificação das ferramentas e técnicas utilizadas, através de uma pesquisa exaustiva na literatura, para o REN em língua portuguesa.

A segunda etapa visa a escolha e posterior utilização das técnicas e ferramentas identificadas na primeira etapa para o REN em relatórios de inteligência financeira (RIF). É importante ressaltar que as entidades mais importantes a serem identificadas no RIF são: pessoas (físicas e jurídicas), valores e datas.

As etapas do processo de REN seguem os algoritmos e técnicas utilizadas para tal finalidade, sendo as mais utilizadas as técnicas de aprendizado de máquina supervisionados, dentre elas citamos o SVM, Redes Bayesianas e redes neurais. Uma das etapas é o treinamento realizados pelos algoritmos, para tanto, é necessário utilizar bases de dados com marcações de entidades para ajudar na identificação automática dessas entidades. Portanto, é fundamental encontrarmos, na literatura, as principais bases de dados anotadas em português.

Será necessária uma etapa intermediária para a conversão dos textos em PDF em texto puro para facilitar a manipulação dos dados.

As métricas utilizadas e que devem aparecer nos resultados são: i) a quantidade de palavras marcadas corretamente como entidade, ii) a quantidade de palavras marcadas indevidamente como entidade e iii) a quantidade de entidades omitidas.

4 RESULTADOS

Abordamos a primeira etapa do nosso processo de revisão da literatura com uma revisão sistemática com o objetivo principal de identificação das principais ferramentas, técnicas e algoritmos utilizados para o reconhecimento de entidades nomeadas em textos em português. Esta etapa está em processo de publicação.

Identificamos na literatura 4 principais ferramentas de análise de textos que fazem o REN para o português, são elas: Stanford CoreNLP[1], OpenNLP[2], spaCy[3] e NLTK[4]. Além disso, identificamos, também, na literatura as principais bases de dados anotadas em português, que são: HAREM[5], BOSQUE[6], FLORESTA[7] e AMAZÔNIA[8].

Identificadas as principais bases de dados anotadas em português e as principais ferramentas utilizadas para o REN em português, procuramos por dados na literatura que nos permitisse escolher qual ferramenta e qual base de dados utilizar para o nosso objetivo final: o reconhecimento de entidades nomeadas em relatórios de inteligência financeira. Optamos, portanto, por utilizar a ferramenta spaCy e a base de dados HAREM por dois motivos principais: os resultados na literatura

1 Encontrada em https://nlp.stanford.edu/software/ acesso em 29 de setembro de 2018.
2 Encontrada em http://opennlp.apache.org/ acesso em 29 de setembro de 2018.
3 Encontrada em https://spacy.io/ acesso em 29 de setembro de 2018.
4 Encontrada em https://www.nltk.org/ acesso em 29 de setembro de 2018.
5 Encontrada em http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip acesso em 25 de setembro de 2018.
6 Encontrado em https://www.linguateca.pt/Floresta/ficheiros/Bosque_CP_7.5_cgde_22032016.conll.gz acesso em 29 de setembro de 2018.
7 Encontrado em https://www.linguateca.pt/Floresta/ficheiros/FlorestaVirgem_CP.conll.gz acesso em 29 de setembro de 2018.
8 Encontrado em https://www.linguateca.pt/Floresta/ficheiros/amazonia.conll.gz acesso em 29 de setembro de 2018.

mostram o spaCy como uma boa ferramenta de REN e o HAREM é uma das bases de dados mais completas e anotadas para o português.

Alguns problemas foram superados e os relatamos para fins de pesquisa: a última versão do HAREM não é compatível com a última versão do spaCy (2.0.12), isto nos levou a desenvolver um script para a conversão dos dados para o spaCy (disponível após o blind review). O spaCy tem as entidades/categorias anotadas em inglês e o HAREM está anotado em português, portanto, fizemos a conversão das categorias no HAREM.

Para a conversão dos RIF do formato PDF para texto, utilizamos a biblioteca Pdfminer.

Os nossos testes foram baseados nos seguintes modelos de treinamentos:

- Modelo 1 - modelo original disponibilizado no SpaCy, versão utilizada 2.0.0;
- Modelo 2 - modelo gerado com base no original, acrescido do treino de 500 iterações dos 'dados de treino' e 20 iterações nos dados dos arquivos de Localidades;
- Modelo 3 - modelo gerado com base no original, acrescido do treino de 20 iterações dos 'dados de treino';
- Modelo 4 - modelo em branco, acrescido do treino de 20 iterações dos 'dados de treino';
- Modelo 5 - novo modelo do HAREM, acrescido de 2 iterações com os dados de treinamento;
- Modelo 6 - novo modelo do HAREM, acrescido de 110 iterações com os dados de treino.
- Modelo 7 - modelo do spaCy com o treinamento junto ao framework Rasa MLU[9]

Segundo Pires (2017), os melhores resultados com o spaCy e o HAREM, para a classificação de entidades nomeadas como notícias em uma base de dados chamada SIGARRA, foram obtidos com 110 iterações de treinamento. Repetimos os mesmos treinamentos, entretanto, é importante ressaltar que as versões, da ferramenta e da base anotada, usadas no trabalho são diferentes.

Todos os RIF usados no treinamento tinham apenas 182 entidades marcadas, sendo 80 do tipo Pessoa (física ou jurídica), 38 do tipo Moeda, 14 do tipo Localidade, 38 do tipo Documento e apenas 12 do tipo Data.

Mostramos os resultados na Tabela 1, após os treinamentos, aplicamos o REN para um dos RIF.

Tabela 1: Resultados da Identificação.

Modelo	Precisão	Recall	f-score
1	1.041	1.587	1.257
2	0	0	0
3	27.586	38.095	32.000
4	14.754	28.571	19.459
5	8.641	11.111	9.722
6	27.272	33.333	30

Fonte: Elaborado pelo autor.

9 Encontrado em https://rasa.com/docs/nlu/. Acessado em 01 de outubro de 2018.

O modelo 7 foi testado em separado, pois não tínhamos total domínio do framework Rasa NLU, pois ele é utilizado em Chatbot para REN. O teste foi realizado a partir do modelo do spaCy e o Rasa NLU como treinamento em cima dos dados do RIF. Neste treinamento, conseguimos identificar 162 entidades, com uma precisão de 60,98%.

A precisão de todos os testes é muito baixa, entretanto, o Modelo 7 chega muito próximo aos valores encontrados por Fonseca et al (2017).

Sabemos que estes primeiros testes são importantes para reconhecermos o nosso corpus e, também, para entendermos os pontos de melhoria dos modelos e dos treinos. Esses resultados estão muito aquém dos resultados encontrados na literatura, seja para a língua portuguesa, seja para a língua inglesa. Cabe ressaltar que o Modelo 3 utilizou os mesmos dados usados para treinamento e validação dos resultados, por isso ele chegou a um resultado próximo ao Modelo 6.

5 CONSIDERAÇÕES FINAIS

Compreendemos que a proposta de automatização de detecção de entidades nomeadas e seus vínculos em RIF é viável utilizando-se as ferramentas e bases de dados existentes. Entretanto, ainda não temos bons resultados com eles, visto que a identificação correta dessas entidades ainda não superou os 90% de acurácia, o que é um grande problema para a análise dos RIF por parte das instituições envolvidas, como é o caso da Polícia Federal. Os modelos treinados, apesar de um melhor desempenho, também não são satisfatórios. Sabemos que é necessário identificar mais entidades em outros RIF para aumentar a qualidade do treinamento. Isto deve levar ao REN com maior precisão, em especial as entidades do tipo Pessoa.

Uma outra proposta para aumentar a assertividade é a criação de expressões regulares e heurísticas que envolvam a identificação de entidades do tipo moeda, localização, data, CPF e CNPJ. Desta maneira, poderemos focar no reconhecimento de entidades do tipo pessoa física e jurídica, para, em um segundo momento estudarmos as correferências entre essas entidades.

Com trabalhos futuros, sugerimos a adaptação das outras bases de dados anotadas (FLORESTA, AMAZÔNIA e BOSQUE) para serem utilizadas como treinamento das principais ferramentas de NER: Stanford CoreNLP, OpenNLP, spaCy e NLTK. Após os testes com todas as bases e ferramentas, buscaremos melhorias nos algoritmos para conseguirmos uma acurácia acima dos 95%, conforme já é apresentado na literatura para o idioma inglês.

REFERÊNCIAS

ARAS, V. Sistema nacional de combate à lavagem de dinheiro e de recuperação de ativos. Revista Jus Navigandi, no 1411, Teresina, 2007.

BRASIL. Lei da Lavagem de Dinheiro, 1998. Disponível em: <http://www.planalto.gov.br/ ccivil_03/ Leis/L9613.htm. Acesso em jun 2018.

CALLEGARI, A. L. Imputação Objetiva. Lavagem de dinheiro e outros temas do Direito Penal. Porto Alegre: Livraria do Advogado, 2001.

COAF. Relatório de Inteligência Financeira - RIF. Disponível em: <http://coaf.fazenda. gov.br/menu/a-inteligencia-financeira/relatorio -de-inteligencia-financeira-rif>. Acesso em: 1 jul. 2018.

FONSECA, Evandro B., CHIELE, Gabriel C., Vieira, Renata e VANIN, Aline A. Reconhecimento de Entidades Nomeadas para o Português Usando o OpenNLP. Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2015), s. pp. Disponível em: <http://www.lbd.dcc.ufmg.br/ colecoes/eniac/2015/011.pdf>. Acesso em: 01 out. 2018.

LUSTOSA, D. S. de M. Aspectos gerais do crime de lavagem de dinheiro (Lei 9.613/98). Âmbito Jurídico, Rio Grande, XII, n. 70, 2009.

MARQUES, N. J. F. O papel do COAF no combate ao crime de lavagem de dinheiro. Conteudo Juridico, Brasilia-DF, 2014.

POLÍCIA FEDERAL. Manual Prático de Combate à Lavagem de Dinheiro e aos Crimes Financeiros (reservado) Brasília - DF, 2013.

PIRES, André R. O. Named entity extraction from Portuguese web text. Faculdade de Engenharia da Universidade do Porto, 2017.

REPOSITÓRIOS DE DADOS DE PESQUISA:
investigando sua adoção nos institutos brasileiros de pesquisa

RESEARCH DATA REPOSITORIES:
investigating its adoption in the brazilian research institutes

André Luiz França de Madeiro
Universidade Federal da Paraíba, Departamento de Ciência da Informação, andreluizfmadeiro@gmail.com

Resumo:
A ciência atual é baseada no uso e compartilhamento de dados de pesquisa por meio das Tecnologias de Informação e da Comunicação. Neste contexto é possível acessar estes dados em bancos de dados *online* conhecidos como repositórios digitais. Os repositórios de dados são um dos tipos desta categoria. Eles podem estar associados a instituições de pesquisa e destacam-se cada vez mais no cenário nacional e internacional. A pesquisa possui como objetivo geral investigar os repositórios de instituições brasileiras de pesquisa no que diz respeito ao suporte para a gestão de dados de pesquisa. A pesquisa apresentou um percurso metodológico bibliográfico, exploratório e quanti-qualitativo. Após a análise dos repositórios encontrados, percebemos uma incipiente presença de repositórios de dados de pesquisa no contexto das instituições de pesquisa brasileiras. Estima-se que este contexto tende a mudar devido à necessidade global de gerenciar novas produções de dados de pesquisa que aumenta constantemente. Novos repositórios de pesquisa poderão surgir à medida que a comunidade de pesquisadores e as instituições estiverem cientes dos benefícios proporcionados por eles e que o tema possa contribuir para novas investigações científicas.
Palavras-chave: Repositório de dados científicos. Instituições brasileiras de pesquisa. Gestão de dados de pesquisa.

Abstract:
Nowadays science is based on the use and sharing of research data through Information and Communication Technologies. In this context it is possible to access this data in online databases known as digital repositories. Data repositories are one of a kind in this category. They can be associated with research institutions and are becoming more relevant in the national and international scenario. The research general objective is to investigate the repositories of Brazilian research institutions regarding the support of research data management. This research presents a bibliographic, exploratory and quanti- qualitative methodology. After the analysis of the repositories found, we noticed an incipient presence of repositories of research data in the context of Brazilian research institutions. It is estimated that this context tends to change due to the global need to manage new and constantly increasing research data productions. New research repositories may emerge as the research community and institutions are aware of the benefits they provide and the subject can contribute to new scientific research.
Keywords: Research Data Repositories. Brazilian Research Institutions. Research Data Management.

1 INTRODUÇÃO

A ciência atual está contextualizada no âmbito de um fenômeno conhecido como e-*Science*[1]. Segundo Oliveira (2016) este termo representa o quarto paradigma científico e o cientista Jim Gray foi seu precursor. No contexto histórico dos paradigmas da ciência, o primeiro paradigma foi marcado pela experimentação, ocorrido há mais de mil anos. O segundo fundamentou-se na ciência teórica com o surgimento de modelos e generalizações, no século

XVII. O terceiro é representado na ciência computacional com base na simulação de fenômenos complexos. O paradigma atual caracteriza-se pela presença da geração intensiva de grande quantidade de dados que surgem do processo de investigação científica (OLIVEIRA, 2016).

Na ciência contemporânea, os dados de pesquisa não são insumos exclusivamente destinados para os pesquisadores que os produziram. Estes dados também devem beneficiar toda a comunidade científica através de estratégias que possibilitem o seu uso e reuso. Os dados podem estar disponibilizados através de repositórios de dados científicos com o acesso livre de barreiras geográficas e de restrições temporais

A relevância da temática em tela é de interesse dos pesquisadores, curadores de dados, comunidades científicas, órgãos de fomento à pesquisa, editores de periódicos científicos, gestores e executores de programas e projetos de dados de pesquisa, profissionais da computação e da Ciência da Informação em geral.

Devido a crescente expansão e compartilhamento do volume de dados de pesquisa, é necessário o estabelecimento e a consolidação de uma política de gestão de dados científicos, visando a garantia dos mesmos por meio de processos de curadoria, preservação, arquivamento e compartilhamento de coleções de dados de pesquisa, em sustentabilidade contínua e permanente (IBICT, 2016).

O valor dos dados de pesquisa em nossa contemporaneidade motivou o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) a lançar o "Manifesto de Acesso Aberto a Dados da Pesquisa Brasileira para Ciência Cidadã". Esta ação foi efetivada na data de 28 de setembro de 2016, configurando-se como um posicionamento brasileiro diante do movimento mundial ao acesso aberto à informação científica. O objetivo desta iniciativa foi incentivar o acesso livre, abrangente e ilimitado às fontes primárias de dados de pesquisa que tornam possível o uso, reuso, compartilhamento e reprodutibilidade em novos contextos de investigação científica (IBICT, 2016).

Dentre os mais variados tipos de repositórios existentes no Brasil, os associados com as instituições de pesquisa têm presença cada vez mais marcante na Internet. A pesquisa de Sayão e Sales (2016) é um exemplo deste assunto: os referidos pesquisadores trazem informações sobre repositórios de dados de pesquisa, sobre sua definição, objetivos, benefícios, apresentação de fluxos de gestão de dados de pesquisa em repositório de dados, tipos de repositórios de dados de pesquisa, valor e permanência de dados e análise dos repositórios *Dryad* e *R3DATA*.

Repositórios de dados associados a instituições de pesquisa são novas entidades que despontam na seara da Tecnologia da Informação. Eles favorecem a propagação das informações científicas e tecnológicas. De acordo com Sayão e Sales (2015, p.53):

1 O termo *data intensive science* é equivalente ao termo e-*Science*. Doravante, adotaremos o termo e-*Science*.

> Os repositórios de dados de pesquisa são banco de dados que recebem, gerenciam e disseminam dados e conjuntos de dados de pesquisa. Eles constituem o lugar mais apropriado para que seus dados sejam preservados e possam ser recuperados, acessados e citados por outros pesquisadores, ou seja, tenham visibilidade em escala mundial.

2 OBJETIVOS

No que se refere a problematização, esta pesquisa questiona de que forma se configuram os repositórios de pesquisa das instituições brasileiras de pesquisa no que diz respeito a gestão de dados de pesquisa.

A pesquisa possui como objetivo geral investigar os repositórios de instituições brasileiras de pesquisa no que diz respeito ao suporte para a gestão de dados de pesquisa. Quanto aos objetivos específicos elencamos os seguintes: identificar os repositórios de dados de institutos brasileiros de pesquisa; identificar a tipologia dos repositórios em institutos brasileiros de pesquisa; levantar características específicas dos repositórios de dados de institutos brasileiros de pesquisa; disponibilizar os dados coletados e o material de apoio usado na pesquisa no repositório de dados no repositório *Dataverse*[2] da UFPB.

3 PROCEDIMENTOS METODOLÓGICOS

A trilha metodológica usada para a condução desta investigação baseou-se numa abordagem bibliográfica, de cunho exploratório, quanti-qualitativo.

A busca por textos para a revisão de literatura e para a identificação dos repositórios de dados foi realizada através das ferramentas de busca Google e Google Acadêmico no período compreendido entre agosto de 2017 a julho de 2018. Os descritores selecionados para a pesquisa foram: "repositórios brasileiros"; "repositórios de dados"; "*brazilian repositories*" e "*data repositories*".

Os *sites R3DATA, ROAR* e *IBICT* foram usados como fontes de referência para identificar os repositórios de dados existentes. O processo investigativo foi bastante exaustivo. Todos os repositórios levantados foram acessados para serem identificados com relação a sua tipologia e as características dos objetos digitais ora depositados. Para auxiliar na identificação de repositórios de dados de pesquisa, foram utilizados os estudos de Leite et al. (2012), Pampel et al. (2013) e Sales (2014).

A partir dos repositórios brasileiros identificados, verificou-se a existência de muitos repositórios institucionais. Após uma análise precisa, consultando os tipos de dados presentes, a finalidade da instituição, informações referentes dos repositórios, verificamos que alguns poucos destes repositórios poderiam ser classificados como repositórios de dados de pesquisa.

2 https://dataverse.ufpb.br/

4 RESULTADOS

Foram encontrados seis repositórios de dados em instituições brasileiras de pesquisa. Dentre os repositórios identificados como repositórios de dados, listamos os seguintes: *PPBiodata*, *IBICT Dataverse, carpeDIEN,* Portal GEOINFO de infraestrutura de dados espaciais da EMBRAPA, Portal da Biodiversidade e o repositório SinBiota (Sistema de Informação Ambiental do Programa Biota/Fapesp). Todos apresentam diversidade na tipologia dos dados em e variados formatos.

O repositório de dados científicos *IBICT Dataverse* encontra-se no *site* https://repositoriopesquisas.ibict.br/. Ele apresenta uma tipologia orientada para várias áreas do conhecimento, de acordo com as informações em seu *site*:

> O IBICT está disponibilizando um repositório de dados de pesquisa que cuida da preservação a longo prazo e de boas práticas de arquivamento, para que os pesquisadores possam compartilhar, manter o controle e obter o reconhecimento de seus dados. O repositório suporta o compartilhamento de dados de pesquisa com citação de dados persistentes, permitindo que eles sejam reproduzidos. O Dataverse é um grande repositório aberto de dados de todas as disciplinas, criado pelo Institute for Quantitative Ciências Sociais da Universidade de Harvard (*IBICT Dataverse*, 2018, *online*).

O repositório IBICT usa o *software Dataverse*. O *Dataverse* (2018, *online,* tradução nossa) é definido como "um aplicativo da *Web* de código aberto para compartilhar, preservar, citar, explorar e analisar dados de pesquisa[3]".

O CarpeDIEN, pode ser acessado através do *site* http://carpedien.ien.gov.br/, é o nome do repositório do *IEN* (*Instituto de Energia Nuclear*). Apresenta uma tipologia voltada para "dados e informações em energia nuclear" (*carpeDIEN*, 2018, *online*) e faz uso do *software DSpace*. O DSpace (*DSPACE*, 2018, *online,* tradução nossa) é conhecido como o "software preferido das organizações acadêmicas, sem fins lucrativos e comerciais que criam repositórios digitais abertos[4]".

O PPBiodata pode ser acessado através do *link* https://ppbiodata.inpa.gov.br/metacatui/. Ele faz parte do *PPBio* (*Programa de Pesquisa em Biodiversidade*). *PPBio* foi:

> criado em 2004 com o objetivo de intensificar estudos sobre biodiversidade no Brasil, descentralizar a produção científica dos centros desenvolvidos academicamente, integrar atividades de pesquisa e divulgar os resultados para diferentes finalidades, como gestão ambiental e educação (*PPBio*, 2018, *online*).

Ainda, com relação ao *PPBiodata*, é importante mencionar que o uso faz uso do *metacat*. Este *software* é um "repositório de dados e metadados (documentação sobre dados), que ajuda os cientistas a encontrar, entender e usar efetivamente os conjuntos de dados que gerenciam ou que foram criados por outros[5]" (KNB, 2018, *online,* tradução nossa).

---

3 Fonte original em língua inglesa: *is an open source web application to share, preserve, cite, explore, and analyze research data.* (DATAVERSE, 2018, *online*).

4 Fonte original em língua inglesa: *DSpace is the software of choice for academic, non-profit, and commercial organizations building open digital repositories* (DSpace, 2018, *online*).

5 Fonte original em língua inglesa: *is a repository for data and metadata (documentation about data), that helps scientists find, understand and effectively use the data sets they manage or that have been created by others* (KNB, 2018, *online*).

Costa (2017) menciona o Portal GEOINFO de infraestrutura de dados espaciais da EMBRAPA e o Portal da Biodiversidade. Peixoto (2012) indica o repositório SinBiota (Sistema de Informação Ambiental do Programa Biota/Fapesp.

Conforme Pampel et al. (2013) e Sales (2014), respectivamente, os repositórios encontrados estão de acordo com a tipologia repositórios de dados de pesquisa institucional e repositórios de dados. Verifica-se uma consonância entre a fundamentação teoria e a investigação realizada.

5 CONSIDERAÇÕES FINAIS

A partir da pesquisa realizada, concluímos que o número de repositórios de dados em instituições de pesquisa brasileira ainda é insipiente. Concordamos com Sales (2014, p.62) sobre este contexto quando afirma que: "este tipo de repositório é mais raro no contexto brasileiro". Contudo, a partir do crescimento exponencial na geração de dados e considerando o valor deste recurso no fazer científico, acreditamos que esta situação poderá modificar-se. Novos repositórios de dados de pesquisa poderão surgir, sejam a partir de iniciativas isoladas das instituições ou mesmo através da geração de consórcios.

REFERÊNCIAS

COSTA, M.M. **Diretrizes para uma política de gestão de dados científicos no Brasil**. Brasília: 2017.288p. Tese (Doutorado em Ciências da Informação). Faculdade de Ciência da Informação. Universidade de Brasília. Disponível em:<http://repositorio.unb.br/handle/104 82/24895>Acesso em: 24 abr. 2018.

DATAVERSE. **About**. Disponível em:< https://dataverse.org/about> Acesso em: 03 jun. 2018.

IBICT. IBICT lança Manifesto de Acesso Aberto a Dados da Pesquisa Brasileira para Ciência Cidadã.

Disponível em:<http://www.ibict.br/Sala- de-Imprensa/noticias/2016/ibict-lanca- manifesto-de-a-cesso-aberto-a-dados- da-pesquisa-brasileira-para-ciencia- cidada> Acesso em: 29 jun. 2018.

KNB. **The Knowledge Network for Biocomplexity**. Disponível em:https://knb.ecoinformatics.org/knb/do cs/intro.html#what-s-in-this-guide. Acesso em: 08 out. 2018.

LEITE, F. et al. **Repositórios Institucionais**: boas práticas para a construção de repositórios institucionais da produção científica. Brasília: IBICT, 2012. Disponível em:< http://livroaberto.ibict.br/bitstream/1/703/ 1/Boas%20pr%C3%A1ticas%20para%2 0a%20constru%C3%A7%-C3%A3o%20 de%20reposit%C3%B3rios%20institucio nais%20da%20produ%C3%A7%-C3%A 3o%20cient%C3%ADfica.pdf> Acesso em: 10 jun. 2018.

OLIVEIRA, A. C. S. **Desvendando a autoralidade colaborativa na e- *Science* sob a ótica dos direitos da propriedade intelectual**. João Pessoa: 2016. 297p. Tese (Doutorado em Ciências

da Informação). Centro de Ciências Sociais Aplicadas. Universidade Federal da Paraíba. Disponível em:<http://tede.biblioteca.ufpb.br/handle/tede/8849>. Acesso em: 7 jun 2018.

PAMPEL, H. et al. **Making research data repositories visible**: the re3data.orgRegistry. PLoSOne, San Francisco, v. 8, n.11, 2013. Disponível em:<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078080/>. Acesso em: 19 maio. 2018.

PEIXOTO, N.M. **Arquitetura de informação em repositório de dados científicos**: análise da interface do repositório do PELD: Programa de Pesquisa Ecológicas de Longa Duração. 2012. Porto Alegre: 2012. 78p. Trabalho de Conclusão de Curso (Graduação) Universidade Federal do Rio Grande do Sul. Faculdade de Biblioteconomia e Comunicação. Curso de Biblioteconomia. Disponível em:< https://lume.ufrgs.br/handle/10183/6972 7>. Acesso em: 29 jul. 2018.

SALES, L.F. **Integração semântica de publicações científicas e dados de pesquisa**: proposta de modelo de publicação ampliada para a área de Ciências Nucleares. Rio de Janeiro: 2014. 265p. Tese apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade Federal do Rio de Janeiro/. Disponível em:<http://repositorio.ibict.br/bitstream/1 23456789/874/1/LUANA%20SALES%20D.pdf>. Acesso em: 05 dez. 2017.

SAYÃO, L.F; SALES, L.F. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, Londrina, v. 21, n. 2, p. 90 – 115, maio/ago 2016. Disponível em:<http://www.uel.br/revistas/uel/index. php/informacao/article/view/27939/2012 2>. Acesso em: 18 abr. 2018.

REVOLUÇÃO CIENTÍFICA?

Reflexões para uma teoria crítica da ciência dos dados

SCIENTIFIC REVOLUTION?
Reflections for a Critical Theory of Data Science

Max Melquiades da Silva[1], Simone Cristina Dufloth[2]
(1) UFMG, Belo Horizonte, iemax@bol.com.br.
(2) Fundação João Pinheiro, Belo Horizonte, sduf@uol.com.br.

Resumo:
O trabalho apresenta uma investigação teórico-conceitual que analisa, à luz da noção de paradigma científico na filosofia da ciência, especialmente os estudos epistemológicos pós-popperianos de Thomas Kuhn (1994) e Inre Lakatos (1979), as características do chamado quarto paradigma da ciência (e-Science). O paradigma da e-Science é apresentado por Gray e outros (2009) como um novo padrão intelectual de produção de conhecimento científico baseado no trabalho intensivo com dados tem conseguido consolidar uma ampla comunidade de prática de pesquisa ao redor do mundo. O trabalho evidencia o excesso de pesquisas de estudo de caso, em contraposição a escassos trabalhos teórico-conceituais que explicitem e discutam os fundamentos teóricos e epistemológicos da e-Science.
Palavras-chave: e-Science. Big data; Epistemologia.

Abstract:
The paper presents a theoretical-conceptual research that analyzes, in light of the notion of scientific paradigm in the philosophy of science, especially the post-Popperian epistemological studies of Thomas Kuhn (1994) and Inre Lakatos (1979), the characteristics of the so-called fourth paradigm of science (e-Science). The paradigm of e-Science is presented by Gray et al. (2009) as a new intellectual standard of scientific knowledge production based on data intensive work has managed to consolidate a wide community of research practice around the world. The work evidences the excess of case study research, as opposed to scarce theoretical-conceptual works that explain and discuss the theoretical and epistemological foundations of e-Science.
Keywords: e-Science; Big data; Epistemology.

1 INTRODUÇÃO

Se a história da ciência no século XX é caracterizada pela emergência de um sem número de recortes temáticos e a consolidação de diversas áreas do conhecimento com o status de ciências (SHAPIN, 1996), o fim do século reservava inovações tecnológicas que permitiriam aumentar a escala e a rapidez das transações informacionais – o que em última instância altera a forma e o ritmo do trabalho científico e a interação entre áreas do conhecimento.

A possibilidade de gerar novos dados de pesquisa em um volume sem precedentes é vista por cientistas de dados como uma realidade análoga à descrita pela lei de Moore ao se referir ao aumento progressivo da capacidade de processamento dos computadores com redução de custos: Em intervalos regulares de tempo o volume de dados cresce e o custo diminui: o desafio se desloca da obtenção para o tratamento e análise desses dados (SAYÃO e SALES, 2015). Nessa linha, Gray aponta para o fato de que a evolução em curso estaria transformando a própria dinâmica da ciência e seus paradigmas: "Quase tudo sobre ciência está mudando por causa do impacto da tecnologia da informação. A ciência experimental, teórica e computacional está sendo afetada pelo dilúvio de dados, e um quarto paradigma da ciência, 'intensivo em dados', está emergindo. O objetivo é ter um mundo em que toda a literatura científica esteja on-line, todos os dados científicos estejam on-line e eles interajam uns com os outros" (GRAY, 2009, 29).

Conforme Gray, uma das maiores referências em se tratando de quarto paradigma da ciência, mesmo nas "ciências de poucos dados" (Idem, p. 20) o trabalho do cientista consiste cada vez mais em analisar a informação disponível – com crescente emprego de recursos de software e hardware – em detrimento dos processos tradicionais de coleta. Nesse sentido, uma das mudanças mais significativas na pesquisa contemporânea residiria justamente no surgimento de "um novo padrão intelectual de produção de conhecimento científico - conhecido como o Quarto Paradigma Científico ou e-Science - caracterizado pela produção e uso intensivo de dados e pelo compartilhamento e colaboração baseados em rede de computadores de alto desempenho" (SALES e SAYÃO, 2015).

Mas como se constitui e evolui um paradigma científico? As transformações em curso no sentido de uma crescente ressignificação do dado como unidade de análise e sentido chegam a constituir um novo paradigma? A produção teórico-conceitual no campo da ciência da informação tem acompanhado o progressivo trabalho com *big data*?

2 OBJETIVOS E METODOLOGIA

Este trabalho busca analisar, à luz da noção de paradigma científico na filosofia da ciência, as características do chamado quarto paradigma da ciência. O trabalho apresenta uma investigação exploratória através de revisão bibliográfica. No próximo tópico são apresentadas características acerca do chamado quarto paradigma da ciência como descritas por alguns de seus principais expoentes. Na sequência, analisa-se como surge um paradigma científico segundo a visão de Thomas Khum e Inre Lakatos, dois dos mais destacados nomes do pensamento científico pós-popperiano. Por fim, são discutidas questões acerca da base teórica que sustenta as discussões e trabalhos empíricos no campo da e-Science.

3 O CHAMADO QUARTO PARADIGMA DA CIÊNCIA: HISTÓRIA E CARACTERÍSTICAS

As inovações na pesquisa científica a partir das mudanças no volume de dados e no emprego de estratégias e instrumentos computacionais para sua análise tem sido chamada de e-Science e o primeiro registro histórico dessa nomenclatura é atribuído ao diretor-geral do Escritório de Ciência e Tecnologia do Reino Unido, John Taylor, em um documento de 1999. A utilização do termo, contudo, foi ganhando amplitude no decorrer da década de 2000, com o avanço das tecnologias de computação em nível de software e hardware. O termo, que experimentou algumas variações semânticas, talvez por sua amplitude, faz referência a métodos de obtenção de resultados científicos através da utilização de computação intensiva, paralela ou distribuída, de imensos volumes de dados.

O termo e-Science tem sido mais amplamente utilizado desde então, como "a aplicação da tecnologia da computação para a realização de investigação científica moderna, incluindo a preparação, experimentação, coleta de dados disseminação de resultados, armazenamento em longo prazo e acessibilidade de todos os materiais gerados através do processo científico, tais como modelagem e análise de dados, notebooks de laboratório, conjuntos de dados brutos e ajustados, produção de manuscritos e versões preliminares, pré-impressões, e publicações impressas e ou eletrônicas" (BOHLE, 2013, p. 1 - tradução nossa).

Outros conceitos têm sido empregados para designar a nova área científica: "eScience", "ciência orientada para dados" (*data-driven Science*), "ciberinfraestrutura", "quarto paradigma", "dos dados ao conhecimento". Em todos eles está presente a ideia de que cientistas da computação se unem a pesquisadores de diferentes domínios para o desenvolvimento de novos conceitos e teorias a partir de grandes massas de dados, em um trabalho que envolve hardware, software e processos que possibilitem a captura, armazenamento, análise e visualização dos dados.

A noção de um "quarto paradigma" da ciência ganha impulso a partir da publicação da obra homônima, em 2009, e da contribuição de diversos autores motivados pelo trabalho do cientista norte-americano James Nicholas "Jim" Gray. Gray foi um renomado cientista da computação que atuou em laboratórios científicos da IBM e da Microsoft. Entusiasta da noção de e-Science, Gray desapareceu no mar em 28 de janeiro de 2007 – junto com seu barco – poucos dias após proferir a palestra que, transcrita, abre o livro em questão[1]. Nele Gray apresenta sua visão sobre a evolução dos paradigmas científicos. Ele concebia a "ciência intensiva em dados" ou "e-Science" como um "quarto paradigma" da ciência (empírica, teórica, computacional e agora baseada em dados). A ciência empírica ou experimental caracteriza o primeiro paradigma, que cedeu lugar às generalizações teórico-conceituais resultantes do uso de modelos abstratos. O terceiro paradigma é baseado nas simulações assistidas por computadores ou outros equipamentos tecnológicos, enquanto o quarto paradigma é resultante da exploração de dados capturados ou gerados pela simulação.

1 Gray, um experiente marinheiro, possuía um iate de quarenta pés. Em 28 de janeiro de 2007, ele não retornou de uma curta viagem solo às Ilhas Farallon, a 43 km de San Francisco, para espalhar as cinzas de sua mãe. O tempo estava claro e nenhuma chamada de socorro foi recebida e nem sinais automáticos de emergência detectados. Uma busca de quatro dias pela Guarda Costeira usando aviões, helicópteros e barcos não encontrou nada. Em 16 de fevereiro, esta busca foi suspensa, e uma busca subaquática usando equipamentos sofisticados terminou em 31 de maio. Em 28 de janeiro de 2012 – exatos cinco anos após seu desaparecimento – Gray foi declarado legalmente morto.

Em sua análise, Gray (2009) aponta alguns caminhos que caracterizam ou deveriam caracterizar o quarto paradigma:

- a comunicação científica pode deixar de apresentar apenas resultados finais na forma de publicações, mas passa a ter a possibilidade de também partilhar os próprios dados brutos da pesquisa, ampliando a possibilidade de falseamento e novas pesquisas utilizando os mesmos dados;
- as bibliotecas digitais podem experimentar novas formas de distribuição das publicações em diferentes países, idiomas e mais baixos custos em relação à publicação em revistas tradicionais;
- a revisão por pares pode experimentar novos processos ou estruturas digitais e serviços como wikis podem ser criados online para ampliar a discussão sobre os trabalhos após a publicação;
- a internet está se tornando um sistema orientado a objeto, cada vez mais dependente de ontologias, vocabulários controlados e esquemas conceituais que possibilitem dar mais confiabilidade ao significado das coisas e à interação entre diferentes sistemas;
- bibliotecas de dados digitais podem ser fomentadas, para além de bibliotecas de publicações acabadas.

O paradigma emergente envolveria colaboração de pesquisadores em computação com pesquisadores de outras áreas do conhecimento, além de métodos computacionais sofisticados para lidar com grandes volumes de dados (*Big Data*), alçados agora à condição de objetos intelectuais de primeira grandeza. Computadores, bancos de dados e redes não são vistos e utilizados apenas como ferramentas, mas se tornam uma parte fundamental do processo de descoberta de conhecimento e mesmo de interpretação dos dados. Assim, o novo paradigma agrega a seu instrumental métodos e algoritmos para analisar os dados armazenados nessas bases de grande escala, além de protocolos de comunicação padronizados que possibilitem a interação entre todas essas fontes de dados.

No Brasil, Sayão e Sales (2015) são alguns dos mais entusiasmados pesquisadores engajados na reflexão e prospecção de um modelo e uma infraestrutura para e-Science, especialmente no que diz respeito a novas fontes de pesquisa, arranjos para novos serviços em diversos estágios do processamento dos dados de pesquisa e uma oportunidade para a reinvenção das bibliotecas de pesquisa. Para eles, "o digital não é o antagônico do impresso, como o rolo de papiro não é o antagônico do livro e a tecnologia não é algo estranho à biblioteca" (SALES e SAYÃO, 2015, p. 33). Embora neguem um eventual antagonismo, os autores se mostram entusiastas de uma ciência orientada para dados. Para eles, "a convergência de tecnologias de computação, armazenamento on-line, redes de alto desempenho, somados ao desenvolvimento de instrumentos científicos, escalas, dispositivos experimentais e sensores cada vez mais sofisticados e ao uso intensivo de simulações, desloca a pesquisa científica contemporânea na direção de uma ciência orientada por dados, onde o maior problema não é a escassez, mas sim o excesso de dados e a capacidade de interpretar seus padrões ocultos na forma" (Idem, p. 34).

4 O SURGIMENTO DE NOVOS PARADIGMAS CIENTÍFICOS

Nesta seção, recorreu-se aos estudos epistemológicos pós-popperianos de Thomas Kuhn (1994) e Inre Lakatos (1979), caracterizados por uma estreita conexão com a história da ciência na tentativa de compreensão dos programas de pesquisa científica[2].

2 Esta seção condensa parte da análise epistemológica desenvolvida por ocasião da dissertação de mestrado de um dos autores deste artigo (SILVA, 2008).

Kuhn se debruçou sobre o que ele chamou de *paradigma científico*, termo com o qual buscava designar um conjunto de "conquistas científicas universalmente reconhecidas, que por certo período fornecem um modelo de problemas e soluções aceitáveis aos que praticam em certo campo de pesquisas" (KUHN, 1994, p. 23).

Esse conceito culmina na noção de *programa de pesquisa*. Com efeito, programa de pesquisa é como Kuhn (1994) chama o estudo acadêmico de determinado tema no interior de uma ciência ou conjunto de ciências – no caso de objetos caracterizados pela interdisciplinaridade. Ao redor daquele programa de pesquisa se reúne um grupo de pesquisadores denominado *comunidade científica* (KUHN, 1994). Dentre os membros da comunidade científica, Kuhn (1994) afirma que alguns ganham destaque sobre os demais devido, basicamente, a dois fatores: a ordem cronológica de publicação dos trabalhos – estudos precursores que enfatizam um aspecto particular do programa de pesquisa e que ganham adeptos podem fazer originar uma nova comunidade científica; outro fator é a originalidade, associada ao grau de aceitação dos trabalhos por parte dos demais membros da comunidade. Na teoria kuhniana a comunidade científica está caracterizada pela prática do que ele chama de *ciência normal*, que é a "pesquisa estavelmente baseada em um ou mais resultados alcançados pela ciência do passado, aos quais uma comunidade científica particular, por certo período de tempo, reconhece a capacidade de constituir o fundamento de sua práxis ulterior" (KUHN, 1994, p. 68). A ciência normal consiste, então, na tentativa de realização das *promessas* do paradigma de pesquisa, aprofundando as pesquisas originárias e confrontando as formulações com a realidade, articulando novos conceitos e sugerindo novas ferramentas com as quais provê o paradigma. A teoria de Kuhn sugere que o trabalho do cientista normal orbita ao redor das noções basilares do paradigma e os eventuais problemas teóricos e práticos que encontrar tendem a significar antes uma deficiência *do pesquisador* do que das construções originárias *do paradigma*. Isso se deve à noção de ciência normal, que consiste no estado de uma ciência na qual suas pesquisas e seus resultados são previsíveis e os métodos e resultados das pesquisas corroboram os resultados esperados da pesquisa. Entretanto, podem ocorrer crises, por meio de anomalias sucessivas que não se conformam com os resultados esperados. Se a quantidade desses "problemas" cresce demasiadamente, inicia-se um período de crise do paradigma em que se começa a questionar a validade de suas concepções iniciais: é a fase da *ciência extraordinária*, em que se colocam os dogmas em questão e cresce o movimento dos pesquisadores 'reformistas'. Esse movimento pode desencadear uma *revolução científica*, momento em que um novo paradigma sobressai em relação ao anterior por meio da "conversão" dos cientistas às promessas do novo paradigma. Kuhn (1994) afirma que essa conversão pode se dar por diversas razões: por convencimento, pela sedução que as novas ideias proporcionam (caráter emocional) por familiaridade, nacionalidade e até mesmo por razões estéticas.

A explicação oferecida por Kuhn ao progresso da ciência evidencia um elemento 'problemático' nesta. A ciência de que o autor fala já não é aquela catedrática da 'objetividade científica', da imparcialidade metodológica e das certezas conceituais: é uma ciência que pressupõe escolhas às vezes arbitrárias, cujos pesquisadores não podem ignorar assaltos da dimensão emocional e precisam endossar os dogmas que a sustentam, sob pena de 'perder seu chão', as bases teóricas de sua produção.

A mesma postura crítica quanto aos pressupostos metodológicos frequentemente empregados na atividade científica é encontrada na obra de Imre Lakatos (LAKATOS, 1979), que

também estuda a temática dos programas de pesquisa e apresenta pontos de convergência e de divergência com o trabalho de Kuhn. Sobre as teorias científicas, Lakatos (1979) argumenta que é componente essencial ao progresso da ciência a concorrência entre programas de pesquisa antagônicos. A partir do estudo da crítica à teoria de Kuhn, Lakatos deu um importante passo na filosofia da ciência ao formular a tese segundo a qual os sistemas teoréticos são compostos por um núcleo forte portador de características que o definem como *programa*, circundado por um cinturão de suporte, constituído de hipóteses e suposições auxiliares que poderiam ser refutadas sem prejuízo para o programa como um todo. Quando submetido a testes, o núcleo forte da teoria que possui o papel de "pressuposto básico" não é afetado, de modo que é o cinturão, e não a teoria central em si, que passa pelo processo de averiguação. Desta forma, uma anomalia em qualquer teste afeta unicamente o cinturão de proteção e jamais a teoria central (SILVA, 2008). Esse núcleo – por decisão metodológica dos pesquisadores – é considerado *não falseável*, isto é, geralmente não se contestam no interior do programa as ideias contidas no seu núcleo fundamental porque sua destruição implicaria problemas para o programa como um todo. A teoria de suporte que acaba sendo refutada pode ser substituída por outra, mas para Lakatos o caso mais comum é o de reforçar a cinta de proteção do núcleo com teorias *ad hoc*, ou seja, teorias auxiliares não possuem propósito algum a não ser o de sustentar o núcleo forte.

Embora sedutora, a teoria de Lakatos não escapou de importantes críticas. Ao extremo, pode-se cogitar que *a priori* nenhum programa de pesquisa poderia ser refutado, uma vez que seu núcleo não-falseável poderia ser mínimo a ponto de conter somente princípios amplamente aceitos, e suas teorias de suporte – que efetivamente desenvolvem as ideias do programa – podem ser pragmaticamente descartadas conforme tenham dificuldade de resistir a novas críticas. Lakatos admitia a existência dessa possibilidade mas asseverava que, para um programa científico, o acúmulo de teorias *ad hoc* é desconfortável. Com efeito, este acúmulo leva a comunidade científica a aderir a determinada teoria em detrimento de outra. A nova teoria nem sequer precisa responder a todas as questões que a anterior respondia, bastando que ela responda um número satisfatório de problemas (dentre esses, alguns que a anterior não conseguia responder) e não ter um número grande de fundamentações *ad hoc* (SILVA, 2008).

5 CONSIDERAÇÕES FINAIS

As breves considerações sobre como Khum e Lakatos entendem a formação de um paradigma científico possibilitam lançar luzes sobre o atual momento histórico. Primeiramente, independentemente da designação que demos é inegável o fenômeno do trabalho com grandes dados na produção científica e seus resultados, cujos exemplos mais notórios são o projeto do genoma humano e os aceleradores de partículas que produzem uma infinidade de dados impossíveis de se processar e analisar sem um suporte computacional. O surgimento de uma comunidade de pesquisa trabalhando com grandes dados é, portanto, um fenômeno do mundo da *práxis* que convive com uma ainda incipiente produção teórico-conceitual que a fundamente. Uma rápida pesquisa em portais de periódicos demonstra que há muito mais pesquisa sobre casos de implementação de e-Science que estudos predominantemente acerca de sua epistemologia, seus fundamentos conceituais ou os

aspectos éticos e políticos sem os quais o pensamento científico não se completa – permanece no campo da técnica.

Por sua vez, tentativas de refutação da e-Science não são raras, sendo Martin Frické um de seus maiores críticos. Frické (2015) é cético em relação à possibilidade de o *big data* produzir uma lógica de descobertas científicas universais ou ferramentas de inferência e refutação que torne a e-Science paradigmática. Para ele, a lógica indutiva por trás do trabalho com dados é falível e contaminada pela teoria: "Os instrumentos são construídos ou adotados à luz do que sabemos ou das teorias que temos", uma vez que "dados não falam (…); é necessário uma enorme quantidade de conhecimento básico ou suposições ou pesquisa prévia" (FRICKÉ, 2015, pp. 5-6).

Tal como vem se configurando, o fenômeno da e-Science, a despeito da *resistência* sofrida, vem diluindo ainda mais algumas tênues fronteiras entre várias ciências – queremos destacar as ciências da informação e da computação – sem chegar a constituir apenas uma disciplina autônoma (e-Science), mas uma concepção de ciência e de fazer científico distinta do que se via até então. Talvez não valha a pena entrar em uma discussão se é adequado ou não chamar esse movimento de um novo paradigma científico. O mundo e a natureza fornecem 'resistência' a nossas conceituações arbitrárias sob a forma de anomalias, situações em que fica claro que algo está errado com as estruturas atribuídas ao mundo por nossas construções conceituais aleatórias. Nesse sentido, o tempo e a produção teórico-conceitual da e-Science em diálogo com a tradição científica poderão atestar seu alcance e falsear seus pressupostos, embora seja forçoso reconhecer a urgência de mais fundamentos (núcleo fundamental e teorias *ad hoc*) para que se possa mais confortavelmente chama-la de um novo paradigma científico. A esse respeito, embora sem mencionar especificamente a e-Science, refletia Boaventura Santos há 30 anos: "Depois da euforia cientista do século XIX e da consequente aversão à reflexão filosófica, bem simbolizada pelo positivismo, chegamos a finais do século XX possuídos pelo desejo quase desesperado de complementarmos o conhecimento das coisas com o conhecimento do conhecimento das coisas" (SANTOS, 1988, p. 57). Precisamos conhecer mais como a e-Science quer conhecer o mundo.

REFERÊNCIAS

BOHLE, Shannon. (2013, 12 June). What is E-science and How Should it be Managed? **Scientific and Medical Libraries. Scilogs. Nature and Spektrum der Wissenshaft**. Disponível em: http://www.scilogs.com/scientific_and_medical_libraries/what-is-e-science-and-how-should-it-be--managed/. Acesso em setembro de 2018.

FRICKÉ, Martin. Big data and its epistemology. **Journal of the Association for Information Science and Technology**. 66, 4, p. 651-661 11 p., 2015.

GRAY, Jim. **Jim Gray on eScience**: A Transformed Scientific Method. Edited by Tony Hey, Stewart Tansley, and Krist in Tolle. In: Hey, T., Tansley, S. and Tolle, K. (Ed.). The Fourth Paradigm: data-intensive scientific discovery. Redmond, WA, USA: Microsoft Research, 2009.

KUHN, Thomas. **A estrutura das revoluções científicas**. São Paulo: Perspectiva, 1994.

LAKATOS, Inre; MUSGRAVE, Alan (Org.). **A crítica e o desenvolvimento do conhecimento**: Quarto volume das atas do colóquio internacional sobre Filosofia da Ciência, realizado em Londres em 1965. Trad.: Octavio Mendes Cajado. São Paulo: Cultrix e EDUSP, 1979.

SALES, Luana Farias; SAYÃO, Luís Fernando. Há Futuro para as Bibliotecas de Pesquisa no Ambiente de eScience? **Informação & Tecnologia**: Marília/João Pessoa, 2(1): 30-52, jan./jul., 2015.

SANTOS, Boaventura de Sousa. Um discurso sobre as ciências na transição para uma ciência pós-moderna. **Estudos avançados.** Vol.2 no.2 São Paulo May/Aug. 1988. http://dx.doi.org/10.1590/S0103-40141988000200007. Acesso em setembro de 2018.

SHAPIN, Steven. **The Scientific Revolution**. Chicago: The University of Chicago Press, 1996.

SILVA, Max Melquíades da. A gestão do conhecimento entre as propostas teórico-conceituais e a efetividade de sua implementação na administração pública: problemas e possibilidades. 2008. 175 f. **Dissertação** (Mestrado em Administração Pública) - Fundação João Pinheiro, Belo Horizonte, 2008.

SUPORTE À GESTÃO DE DADOS DE PESQUISA NAS BIBLIOTECAS:
identificação de serviços a partir da literatura

SUPPORT FOR RESEARCH DATA MANAGEMENT IN LIBRARIES:
identification of services from the literature

Fernanda Gomes Almeida[1], Marcos de Souza[2]
(1) Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte, MG, usernanda@gmail.com
(2) Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte, MG, marcosdesouza82@gmail.com

Resumo:
Artigo de natureza exploratória que objetiva inventariar, por meio da literatura, os serviços para suporte à gestão de dados científicos realizados em bibliotecas. Pretende ainda propor uma categorização e classificação dos serviços conforme as categorias propostas. A identificação dos artigos que compõem o *corpus* desta pesquisa foi realizada por meio de buscas em bases de dados, Google Acadêmico e referências citadas. Os artigos foram lidos, codificados e categorizados com o auxílio do *software* NVivo, sendo as categorias criadas a partir da revisão de literatura. Foram obtidas cinco categorias de serviços nos 58 artigos analisados: Infraestrutura, Técnico, Suporte, Educacional e Outros. Os serviços mais presentes nos artigos são os das categorias Educacional e Suporte, representando uma extensão dos serviços já realizados pelas bibliotecas. Conclui-se que a classificação proposta carece de reanálise e reagrupamento de serviços, pois é provável que a partir da literatura e de um melhor entendimento da área surja uma nova hierarquização ou que categorias sejam unificadas ou remanejadas.
Palavras-chave: Dados de pesquisa. Gestão de dados de pesquisa. Bibliotecas. Serviços.

Abstract:
This is an exploratory paper that aims to inventory, through the literature, the services to support for research data management in libraries. It also intends to propose a categorization and classification of services according to the proposed categories. The identification of the articles that compose the corpus of this research was carried out through searches in databases, Google Scholar and references cited. The articles were read, coded and categorized with the aid of the NVivo software, and the categories were created based on the literature review. Five categories of services were obtained in the 58 analyzed articles: Infrastructure, Technical, Support, Educational and Others. The most present services in the articles are those of the Educational and Support categories representing an extension of services already performed by libraries. It is concluded that the proposed classification requires a reanalysis and regrouping of services, since it is probable, that from the literature and from a better understanding of the area a new hierarchy will emerge or that categories will be unified or remanaged.
Keywords: Research data. Research data management. Libraries. Services.

1 INTRODUÇÃO

Atualmente é reconhecido o potencial informacional dos dados de pesquisa por parte de pesquisadores, instituições acadêmicas e agências de fomento. Além desse reconhecimento, uma série de pressões externas como o reuso de dados, os requisitos das agências financiadoras, as boas práticas de pesquisa, a reputação das instituições (ASHLEY, 2012), bem como o crescimento e o desenvolvimento da pesquisa colaborativa, as práticas de pesquisa aberta (*Open Access, Open Science, Open Data*) (CHIWARE; BECKER, 2018) vêm fazendo com que o gerenciamento de dados científicos seja uma preocupação crescente nas universidades (ASHLEY, 2012). Desta forma, o potencial dos dados aliado às pressões externas representam oportunidades para que as bibliotecas ofereçam serviços voltados ao gerenciamento de dados (KRAFT, 2015).

Os serviços de suporte à gestão de dados de pesquisa (SSGDPs) são os "[...] serviços que uma biblioteca oferece aos pesquisadores em relação ao gerenciamento de dados [...]" (TENOPIR; BIRCH; ALLARD, 2012, p. 7, tradução nossa), abordando o ciclo de vida completo dos dados.

Diversos nomes foram sugeridos na literatura para designar basicamente os mesmos tipos de serviços, contudo uma pesquisa não exaustiva apontou que "*research data management services*" e "*research data services*" foram os temos mais utilizados.

As universidades – e as bibliotecas - estão começando a descobrir como a gestão de dados de pesquisa deve ser apoiada em termos de consultoria, treinamentos ou de infraestrutura para armazenamento, compartilhamento e curadoria de dados (COX; PINFIELD, 2014), porém, quais são os serviços que as bibliotecas podem oferecer?

Nos últimos anos diversos artigos têm identificado os variados SSGDPs desenvolvidos e/ou planejados pelas bibliotecas nos Estados Unidos e Europa (SI *et al.*, 2015; TENOPIR *et al.*, 2016, 2017; TENOPIR; BIRCH; ALLARD, 2012; YOON; SCHULTZ, 2017), assim como outros trabalhos têm apresentado os serviços desenvolvidos em instituições (AKERS *et al.*, 2014; GOLDMAN; KAFEL; MARTIN, 2015; HIOM *et al.*, 2015; KNUTH; JOHNSON; HAUSER, 2015; PERRIER *et al.*, 2018; SEARLE *et al.*, 2015; SESARTIC; TÖWE, 2016; TENOPIR *et al.*, 2016, 2017; YOON; SCHULTZ, 2017).

Este trabalho tem como objetivo inventariar, por meio de revisão de literatura, os serviços de suporte à gestão de dados desenvolvidos em bibliotecas. Além disso, pretende propor uma categorização de serviços, bem como classificar esses serviços conforme as categorias propostas.

3 REFERENCIAL TEÓRICO

Recentemente foi percebido que a gestão dos dados de pesquisa constitui o aspecto mais desafiador para as instituições (PRYOR, 2014). Para as bibliotecas, a gestão de dados representa uma oportunidade para que desempenhem um papel ainda mais ativo no processo de pesquisa. Em primeiro lugar, as bibliotecas acadêmicas podem fornecer serviços de consultoria relacionados ao gerenciamento e conservação de dados de pesquisa. Segundo, as bibliotecas podem fornecer a infraestrutura para armazenamento e curadoria de dados. Em terceiro lugar, as bibliotecas acadêmicas podem apoiar os bibliotecários a se tornarem membros ativos em equipes de pesquisa e de propostas

de concessão como consultores em curadoria de dados (TENOPIR; BIRCH; ALLARD, 2012, p.41, tradução nossa).

Nos últimos anos, autores têm proposto níveis, classificações ou categorias de serviços. Algumas dessas são apresentadas na próxima seção.

3.1 Trabalhos relacionados

Geraci, Humphrey e Jacobs (2012) propõem a classificação de serviços em três níveis principais: serviços de coleção – criação e manutenção de coleções; serviços de referência em dados ou curadoria de dados; e serviços de computação – armazenamento, entrega e uso de dados.

Reznik-Zellen, Adamick e McGinty (2012) propõem três níveis de serviço: educação - bibliotecas educam sua comunidade sobre gestão de dados. Este é um nível mais básico de serviços; consulta - suporte sobre uma variedade de questões relevantes para a gestão de dados de pesquisa e; infraestrutura – fornecimento de infraestrutura para gerenciamento e curadoria de dados.

Hiom *et al* (2015) classificam os serviços em: assistência ao gerenciamento de dados; *advocacy* e treinamento; armazenamento e suporte à publicação e; presença na web, mídia social e métricas.

Si *et al* (2015) dividem os serviços em seis aspectos: introdução aos dados de pesquisa; diretrizes do plano de gerenciamento de dados; curadoria de dados e serviços de armazenamento; treinamento em gerenciamento de dados; referência em gerenciamento de dados e; recomendação de recursos.

Tenopir, Birch e Allard (2012) afirmam que os serviços podem ser do tipo informativo/consultivo ou serviços técnicos. Segundo esses autores, os serviços informativos/consultivos são os mais comumente presentes nas bibliotecas.

Yoon e Schultz (2017) examinaram *websites* de bibliotecas acadêmicas americanas e classificaram os seus serviços em quatro áreas: serviço, informação, educação e rede (*links* para recursos externos).

Contudo, a classificação varia bastante, dependendo do entendimento dos autores.

4 PROCEDIMENTOS METODOLÓGICOS

Esta é uma pesquisa de natureza exploratória realizada por meio de levantamento de artigos publicados em periódicos, cuja temática é o suporte à gestão de dados de pesquisa realizado nas bibliotecas. Os artigos foram identificados a partir de pesquisas em bases de dados como a *Library and Information Science Abstracts (LISA*), Library, Information Science & Technology Abstracts (LISTA), *Information Science & Technology Abstracts (ISTA)*, Google Acadêmico e referências citadas.

A busca retornou 160 resultados, contudo, este trabalho apresenta o inventário e categorização obtida a partir de serviços mencionados em 58 (36,25%) artigos. Assim, este trabalho retrata os primeiros resultados da pesquisa, uma vez que, por questão de tempo, não foi possível realizar o inventário e classificação de todos os 160 artigos encontrados na pesquisa.

Os 160 artigos foram importados para o *software* de pesquisa qualitativa NVivo, da QSR *International*. Desses, 58 artigos foram lidos, codificados e categorizados. A seleção desses 58 artigos que compõem o *corpus* desta pesquisa foi realizada ao acaso, sem intenção prévia.

Para a categorização dos serviços, foram criados nós - categorias de serviços -, estruturados em cinco temas principais – Infraestrutura, Técnico, Suporte, Educacional e Outros.

A categoria "Infraestrutura" é formada pela infraestrutura técnica – como os repositórios e ferramentas ou plataformas para assistência à pesquisa e à gestão de dados - e as políticas institucionais.

Na categoria "Técnico" estão as tarefas técnicas realizadas pelas bibliotecas como a documentação de dados e metadados, depósito de dados, preservação, garantia de acesso, publicação, dentre outros.

A categoria "Suporte" é formada pelos serviços que visam à informação, orientação, aconselhamento e consultoria sobre os mais variados assuntos referentes à gestão e curadoria de dados, como a elaboração de planos de gestão de dados, padrões de metadados, acesso aberto aos dados, melhores práticas de pesquisa, e outros.

A categoria "Educacional" abrange os serviços voltados para a formação em gestão e curadoria de dados como treinamentos, cursos, eventos, elaboração de guias, *templates*, *websites*, *data literacy*, etc.

Nos serviços classificados na categoria "Outros" estão a auditoria para identificação das práticas e necessidades dos pesquisadores; o desenvolvimento de políticas, padrões, diretrizes; a divulgação e promoção de serviços; a identificação de colaboradores, participação em grupos de pesquisa; o planejamento, implementação e avaliação de serviços; etc.

As categorias e subcategorias de serviços (nós e subnós) seguem uma estrutura hierárquica e foram criados com base no referencial teórico da pesquisa e a partir dos temas que emergiram dos artigos analisados.

Após a categorização inicial dos serviços, foi realizado um rearranjo das categorias, hierarquização de termos, uniformização de nomes e união das categorias com assuntos que apresentam afinidades temáticas.

5 RESULTADOS

Os serviços mencionados nos 58 artigos foram classificados em cinco categorias principais, conforme a Figura 1.

Figura 1: Categorias de serviços

Name	Files	References
Categorias de serviços	58	1065
1 - INFRAESTRUTURA	21	53
2 - TÉCNICO	39	214
3 - SUPORTE	42	378
4 - EDUCACIONAL	46	279
5 - OUTRO	35	141

Fonte: Elaborado pelos autores

Dos 58 artigos analisados, 21 apresentam serviços classificados na categoria Infraestrutura; 39 mencionam serviços da categoria Técnico; 42 apontam serviços agrupados em Suporte; 46 artigos mencionam serviços na categoria Educacional; enquanto 35 artigos tratam de serviços categorizados como Outros. Vale lembrar que um mesmo artigo pode apresentar serviços presentes em mais de uma categoria ao mesmo tempo.

Percebe-se que a maioria dos trabalhos menciona os serviços da categoria "Educacional" (46 artigos ou 79,31%), seguidos pelos serviços da categoria "Suporte" (42 artigos ou 72,41%), conforme apresentado na Figura 1. Assim, os serviços educacionais e de suporte são os mais frequentes, representando uma extensão dos serviços já realizados pelas bibliotecas, ampliados para o trabalho com dados científicos.

Estes resultados estão de acordo com os trabalhos realizados por Tenopir, Birch e Allard (2012) e Tenopir et al (2017) que identificaram que os serviços informativos/consultivos são os mais realizados pelas bibliotecas; e Reznik-Zellen, Adamick e McGinty (2012) que afirmam que os serviços do nível educacional são os mais básicos. Segundo esses autores, mesmo as bibliotecas que estão fazendo pouco para apoiar a infraestrutura de dados de pesquisa em suas instituições ainda se envolvem em algum tipo de educação.

As categorias e subcategorias completas são apresentadas no Apêndice A.

6 CONSIDERAÇÕES FINAIS

Este trabalho faz parte de uma pesquisa ainda em andamento, apresentando resultados preliminares. Assim, não foi possível realizar a categorização dos serviços a partir de todos os artigos recuperados na busca realizada para a pesquisa, contudo o *corpus* utilizado nesta exploração permitiu chegar a um inventário satisfatório de serviços, uma vez que a cada nova inclusão de artigo, notou-se que havia pouca "novidade" em relação ao acréscimo de novos serviços.

Embora não seja possível listar aqui todos os serviços voltados para a gestão de dados de pesquisa realizados pelas bibliotecas, considera-se que este trabalho cumpriu seu objetivo de inventariá-los, bem como criar uma categoria de serviços e agrupá-los conforme a categorização proposta.

Uma visão geral desses serviços e categorias pode ser visualizada no Apêndice A. Em trabalho futuro, pretende-se descrever minuciosamente cada uma das categorias, assim como todos os serviços que a compõem.

Acredita-se, contudo, que a classificação proposta carece de uma reanálise e reagrupamento de serviços e categorias. A partir de um melhor entendimento da área é provável que surja uma nova hierarquização de categorias ou que elas possam ser unificadas ou remanejadas.

REFERÊNCIAS

AKERS, K. G. *et al.* Building Support for Research Data Management: Biographies of Eight Research Universities. *International Journal of Digital Curation*, v. 9, n. 2, p. 171–191, 2014. Disponível em: <http://www.ijdc.net/index.php/ijdc/article/view/327>. Acesso em: 27 ago. 2018.

ASHLEY, K. Research data and libraries: who does what. *Insights: the UKSG journal*, v. 25, n. 2, p. 155–157, 2012. Disponível em: <http://insights.uksg.org/articles/10.1629/2048-7754.25.2.155>. Acesso em: 21 set. 2018.

CHIWARE, E. R. T.; BECKER, D. A. Research Data Management Services in Southern Africa: a Readiness Survey of Academic and Research Libraries. *African Journal of Library, Archives and Information Science*, v. 28, n. 1, p. 1–16, 2018. Disponível em: <http://search-ebscohost-com.ez27.periodicos.capes.gov.br/login.aspx?direct=true&db=lih&AN=130261946&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid>. Acesso em: 30 jul. 2018.

COX, A. M.; PINFIELD, S. Research data management and libraries: current activities and future priorities. *Journal of Librarianship and Information Science*, v. 46, n. 4, p. 299–316, 28 dez. 2014. Disponível em: <http://lis.sagepub.com/cgi/doi/10.1177/0961000613492542>. Acesso em: 24 ago. 2017.

GERACI, D.; HUMPHREY, C.; JACOBS, J. *Data basics: an introdutory text*. Ann Arbor: Inter-university Consortium for Political and Social Research (ICPSR), 2012. Disponível em: <https://doi.org/10.7939/R3251FK8F>. Acesso em: 14 set. 2018.

GOLDMAN, J.; KAFEL, D.; MARTIN, E. Assessment of Data Management Services at New England Region Resource Libraries. *Journal of eScience Librarianship*, v. 4, n. 1, p. e1068, 16 jul. 2015. Disponível em: <http://escholarship.umassmed.edu/jeslib/vol4/iss1/4/>. Acesso em: 22 ago. 2018.

HIOM, D. *et al.* Research data management at the University of Bristol. *Program: electronic library and information systems*, v. 49, n. 4, p. 475–493, set. 2015. Disponível em: <http://www.emeraldinsight.com/doi/10.1108/PROG-02-2015-0019>. Acesso em: 29 maio 2018.

KNUTH, S. L.; JOHNSON, A. M.; HAUSER, T. Research Data Services at the University of Colorado Boulder. *Bulletin of the Association for Information Science and Technology*, v. 41, n. 6, p. 35–38, ago. 2015. Disponível em: <https://www.asist.org/publications/bulletin/aug-2015/research-data-services/>. Acesso em: 30 jul. 2018.

KRAFT, A. RADAR - A repository for long tail data. 2015, Hanover, DE: IATUL, 2015. Disponível em: <https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2099&context=iatul>. Acesso em: 13 fev. 2018.

PERRIER, L. *et al.* Developing research data management services and support for researchers: a mixed methods study. *Partnership: The Canadian Journal of Library and Information Practice and Research*, v. 13, n. 1, p. 1–24, 2018. Disponível em: <https://journal.lib.uoguelph.ca/index.php/perj/article/view/4115>. Acesso em: 30 jul. 2018.

PRYOR, G. Preface. In: PRYOR, G.; JONES, S.; WHYTE, A. (Org.). . *Delivering research data management services: fundamentals of good practice*. London: Facet Publishing, 2014. p. vii–viii.

REZNIK-ZELLEN, R.; ADAMICK, J.; MCGINTY, S. Tiers of Research Data Support Services. *Journal of eScience Librarianship*, v. 1, n. 1, p. 27–35, 2012. Disponível em: <http://escholarship.umassmed.edu/jeslib/vol1/iss1/5/>. Acesso em: 29 maio 2018.

SEARLE, S. *et al.* Librarians as partners in research data service development at Griffith University. *Program: electronic library and information systems*, v. 49, n. 4, p. 440–460, set. 2015. Disponível em: <http://www.emeraldinsight.com/doi/10.1108/PROG-02-2015-0013>. Acesso em: 6 maio 2018.

SESARTIC, A.; TÖWE, M. Research Data Services at ETH-Bibliothek. *IFLA Journal*, v. 42, n. 4, p. 284–291, 27 dez. 2016. Disponível em: <http://journals.sagepub.com/doi/10.1177/078uopledrsf-drt340035216674971>. Acesso em: 30 jul. 2018.

SI, L. *et al.* Investigation and analysis of research data services in university libraries. *The Electronic Library*, v. 33, n. 3, p. 417–449, jun. 2015. Disponível em: <http://www.emeraldinsight.com/doi/10.1108/EL-07-2013-0130>. Acesso em: 30 jul. 2018.

TENOPIR, C. *et al.* Research Data Services in European Academic Research Libraries. *LIBER QUARTERLY*, v. 27, n. 1, p. 23–44, fev. 2017. Disponível em: <https://www.liberquarterly.eu/articles/10.18352/lq.10180/>. Acesso em: 10 ago. 2017.

TENOPIR, C. *et al.* Research data services in european and north american libraries: Current offerings and plans for the future. *Proceedings of the Association for Information Science and Technology*, v. 53, n. 1, p. 1–6, 2016. Disponível em: <http://doi.wiley.com/10.1002/pra2.2016.14505301129>. Acesso em: 20 ago. 2018.

TENOPIR, C.; BIRCH, B.; ALLARD, S. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*. sl.: Association of College & Research Libraries (ACRL), 2012. Disponível em: <http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf>. Acesso em: 29 jun. 2017. (Whipe Paper).

YOON, A.; SCHULTZ, T. Research Data Management Services in Academic Libraries in the US: A Content Analysis of Libraries' Websites. *College & Research Libraries*, v. 78, n. 7, p. 1–19, nov. 2017. Disponível em: <http://crl.acrl.org/index.php/crl/article/view/16788>. Acesso em: 28 maio 2018.

APÊNDICE A – CATEGORIZAÇÃO DE SERVIÇOS

Categoria	Quantidade de artigos	Referências
1 - INFRAESTRUTURA	21	53
Ferramentas	5	11
Assistência a produtividade	1	3
Mineração e visualização de dados	1	1
Plano de gestão de dados (PGD)	1	1
Trabalho colaborativo	4	6
Plataformas	5	6
Ambiente de pesquisa virtual	1	1
Curadoria de dados - geral	1	1
eResearch	1	1
Gestão de dados - geral	1	2
Publicação de dados	1	1
Políticas	8	13
Repositórios de dados	12	23

Categoria	Quantidade de artigos	Referências
2 - TÉCNICO	39	214
Acesso aos dados	7	9
Arquivamento e preservação de dados	19	30
Catálogo de dados	2	4
Coleta e criação de dados	3	5
Compartilhamento, publicação e depósito de dados	10	18
Concepção e gestão de bases de dados	1	1
Copyright, propriedade intelectual e licenciamento	2	2
Curadoria de dados - geral	8	14
Documentação de dados e metadados	24	42
Gestão da informação de pesquisa	1	1
Gestão de dados - geral	7	9
Mineração, análise e visualização de dados	8	16
Organização, administração, armazenamento e backup	13	20
Qualidade dos dados	2	2
Rastreamento de impacto	3	4
Repositórios de dados	16	30
Uso e reuso de dados	6	7

Categoria	Quantidade de artigos	Referências
3 - SUPORTE	42	378
Acesso aberto	4	5
Arquivamento e preservação de dados	13	19
Coleta e criação de dados	6	9
Compartilhamento, publicação e depósito de dados	17	27
Comunicação científica	1	2
Copyright, propriedade intelectual e licenciamento	15	27
Curadoria de dados - geral	6	9
Design da pesquisa	2	2
Dispositivos móveis	1	1
Documentação de dados e metadados	22	30
eScience - geral	1	1
Financiamento, requisitos dos financiadores e editores	7	9
Fluxos de trabalho de pesquisa	2	3
Gestão de dados - geral	23	35
Melhores práticas para gerenciamento de dados	6	7
Mineração, análise de dados e visualização de dados	8	16
Organização, administração, armazenamento e backup	8	13
Plano de gestão de dados (PGD)	33	62
Políticas	2	2
Qualidade dos dados	1	7
Questões éticas, legais e de consentimento	4	5
Rastreamento de impacto	6	7
Repositórios de dados	8	11
Serviços e sistemas oferecidos	5	7
Software estatístico ou qualitativo	3	5
Tecnologias, infraestruturas e ferramentas	1	2
Uso e reuso de dados	23	55

Categoria	Quantidade de artigos	Referências
5 - OUTRO	35	141
Auditoria	16	30
Identificação das percepções, práticas e necessidades	12	20
Identificação de lacunas nos serviços	2	2
Identificação de recursos e serviços disponíveis	4	4
Desenvolvimento de currículo para gerenciamento de dados	4	4
Desenvolvimento de políticas, diretrizes, serviços e infraestruturas	11	15
Discussão sobre gestão de dados e serviços de dados	5	5
Divulgação, comunicação e promoção dos serviços	13	33
Identificação de colaboradores e estabelecimento parcerias	17	35
Implementação de políticas	1	1
Participação em disciplina	1	1
Participação em grupo de trabalho	3	3
Participação em projeto de pesquisa	6	7
Planejamento, implementação e avaliação de serviços	6	7

Categoria	Artigos	Refs
4 - EDUCACIONAL	46	279
Advocacy	9	17
Documentação de dados e metadados	1	1
Envolvimento da alta administração	1	1
Escience	1	1
Gestão de dados - geral	4	5
Melhores práticas para gestão de dados	2	2
Openness e políticas de acesso aberto	3	4
Plano de gestão de dados (PGD)	2	2
Serviços oferecidos	1	1
Data librarianship	3	6
Data literacy	7	11
Eventos - conferência, seminário, simpósio, palestra	6	9
Guias, templates, tutoriais ou help	18	35
Ajuda com dados ou conjunto de dados	3	3
Compartilhamento, publicação e depósito de dados	1	1
Documentação de dados e metadados	1	1
Gestão de dados - geral	8	11
Organização, administração, armazenamento e backup	1	1
Plano de gestão de dados (PGD)	1	1
Repositórios de dados	2	2
Uso e reuso de dados	4	5
Links	1	9
Externas	1	4
Internos	1	5
Treinamentos, oficinas, cursos e aulas	38	142
Arquivamento e preservação de dados	1	1
Coleta e criação de dados	1	1
Compartilhamento, publicação e depósito de dados	5	8
Curadoria de dados - geral	4	5
Documentação de dados e metadados	4	7
Ferramentas de ponta - geral	1	1
Financiamento, requisitos dos financiadores e editores	3	3
Gestão de dados - geral	25	57

Categoria	Artigos	Refs
4 - EDUCACIONAL - Continuação		
Treinamentos, oficinas, cursos e aulas		
Melhores práticas para gestão de dados	7	12
Metodologia de dados específica por disciplina	1	1
Organização, administração, armazenamento e backup	2	3
Plano de gestão de dados (PGD)	8	12
Privacidade e segurança de dados	1	1
Questões éticas, legais e de consentimento	1	1
Reciclagem para bibliotecários	1	1
Serviços oferecidos	5	7
Software estatístico ou qualitativo	2	2
Uso de conjuntos de dados específicos por disciplina	1	1
Uso de repositórios de dados - busca, upload,	1	1
Uso e reuso de dados	1	2
Website	10	50
Arquivamento e preservação de dados	1	1
Ciclo de vida dos dados	1	1
Coleta e criação de dados	2	3
Compartilhamento, publicação e depósito de dados	4	5
Copyright, propriedade intelectual e licenciamento	2	3
Dados de pesquisa - geral	1	1
Documentação de dados e metadados	4	5
Financiamento, requisitos dos financiadores e editores	1	3
Gestão de dados - geral	8	10
Melhores práticas para gestão de dados	1	1
Organização, administração, armazenamento e backup	2	4
Plano de gestão de dados (PGD)	3	4
Políticas	1	1
Proteção de dados	1	1
Serviços oferecidos	1	1
Treinamento, aconselhamento e suporte	1	1
Uso e reuso de dados	1	5

TECNOLOGIA BLOCKCHAIN:
um Novo Paradigma no Ciclo de Vida dos Dados

BLOCKCHAIN THECNOLOGY:
a New Paradigm in the Data Life Cycle

Jailton Cardoso da Cruz[1]
(1) Instituto Federal de Alagoas, https://www2.ifal.edu.br, jailton@ifal.edu.br

Resumo:
A tecnologia *Blockchain*, também denominada D*istributed Ledger Technology*, é uma inovação desruptiva, ou seja, ela representa um salto inesperado e de grande impacto no que diz respeito a gerenciamento de dados. Criado pela necessidade de um sistema eficiente, econômico, confiável e seguro para conduzir e registrar transações financeiras. Este trabalho apresenta a Tecnologia Blockchain, no contexto do Ciclo de Vida dos Dados, neste sentido, daremos uma ênfase à nova perspectiva de sua utilização para a preservação de dados e as diversas possibilidades de aplicação nas ciências abertas, bem como no ciclo de vida dos dados de pesquisa, para tal, utilizaremos uma estratégia de revisão bibliográfica com foco em estudos referentes à tecnologia Blockchain, bem como o referencial teórico sobre bibliotecas de dados e, a partir daí, fazer uma intersecção sobre o uso da tecnologia Blockchain nas diversas fases do ciclo de vida dos dados científicos com objetivo de apresentar o potencial da tecnologia Blockchain como infraestrutura tecnológica para a gestão de dados científicos. Ao final apresentamos algumas discussões a respeito do tema.
Palavras-chave: Tecnologia Blockchain. Ciências Abertas. Ciclo de Vida dos Dados.

Abstract:
Blockchain technology, also known as Distributed Ledger Technology, is a desruptive innovation, that is, it represents an unexpected and major leap in data management. Created by the need for an efficient, economical, reliable and secure system for conduct and record financial transactions. In this paper we present the Blockchain Technology, in the context of the Data Life Cycle, in this sense, we will emphasize the new perspective of its use for the preservation of data and the diverse possibilities of application in the open sciences, as well as in the life cycle of the To do this, we will use a literature review strategy focused on studies related to Blockchain technology, as well as the theoretical reference on data libraries and, from there, to make an intersection on the use of Blockchain technology in the various phases of the life cycle of scientific data in order to present the potential of Blockchain technology as a technological infrastructure for the management of scientific data. At the end we present some discussions on the subject.
Keywords: Blockchain Technology. Open Science. Data Life Cicle.

1 INTRODUÇÃO

Com o desenvolvimento das tecnologias para o armazenamento, acesso e compartilhamento das informações científicas e, mais recentemente, com as possibilidades que surgiram com o advento da internet, a utilização das novas tecnologias tem sido um desafio para garantir a preservação da produção científica, desde sua origem bem como garantir todo o ciclo de vida dos dados científicos.

Segundo (ARELLANO,2008, p20, apud SOUZA, 2012)

> [...] a preservação digital é uma preocupação dos profissionais da informação, devido principalmente ao uso extensivo que tem sido feito das tecnologias digitais e do registro de informações em soluções cada vez mais sofisticadas e complexas.

Dentre estes desafios referentes à utilização da tecnologia da informação na preservação da produção científica, surge a tecnologia *Blockchain* como um novo paradigma nesse campo. *Blockchain*, conforme (FORMIGONE, 2017, p6),

> Em linhas gerais, pode-se dizer que se trata de um sistema distribuído de base de dados em log, mantido e gerido de forma compartilhada e descentralizada (através de uma rede *peer-to-peer*, P2P), na qual todos os participantes são responsáveis por armazenar e manter a base de dados. A tecnologia foi construída tendo em mente quatro principais características arquiteturais: segurança das operações, descentralização de armazenamento e/ou computação, integridade de dados e imutabilidade de transações".

O relatório *Ways to Use Blockchains in Libraries* (ALMAN, 2017) publicado no site *Blockchains for Information Professional*, que faz parte do projeto de Sue Alman e Sandy Hirish, apresenta alguns cenários em que os profissionais da informação podem fazer com o uso da tecnologia no *Blockchain*:

- Criar um sistema de metadados aprimorado para bibliotecas e *data centers;*
- Controlar os direitos de primeira venda digital e propriedade;
- Conectar-se a uma rede de bibliotecas/universidades;
- Apoiar coleções baseadas na comunidade;
- Hospedar compartilhamento digital ponto a ponto;
- Facilitar a parceria entre centros e organizações.

2 OBJETIVOS

Este artigo tem por objetivo apresentar uma visão geral da tecnologia *Blockchain*, e sua aplicação no campo das ciências de dados.

3 PROCEDIMENTOS METODOLÓGICOS

Quanto ao percurso metodológico, a pesquisa foi desenvolvida uma pesquisa exploratória com objetivo de proporcionar maior conhecimento da tecnologia. Neste sentido, foi feito de uma

revisão de literatura, onde traçamos as estratégias para desenvolvimento do método de pesquisa a ser utilizado com as seguintes fases:

- Revisão das referências bibliográficas referentes ao tópico Gestão de Dados Científicos na Perspectiva da e-*Science*;
- Estudo da tecnologia *Blockchain* a partir do referencial teórico;
- Estudo das bibliotecas de dados, o ciclo de vida dos dados científicos, Gestão de Dados de Pesquisa (GDP);
- Pesquisa bibliográfica sobre a utilização da tecnologia *Blockchain* nas bibliotecas de dados e nas ciências;
- Analise sobre como a tecnologia *Blockchain* poderia ser aplicada no ciclo de vida dos dados científicos.

Neste sentido desenvolvemos o trabalho que, devido a característica desruptiva da tecnologia *Blockchain* traz uma reflexão sobre seu uso no campo da e-*Science*.

4 A TECNOLOGIA *BLOCKCHAIN*

Instrumentos de confiança, tais como papel moeda, moedas cunhadas, carta de crédito e sistemas bancários surgiram para facilitar a troca de valor e proteger compradores e vendedores. Esses métodos tradicionais de transações conforme (GUPTA, 2017) pode ser caro, em parte porque envolve intermediários que cobram taxas por seus serviços. É claramente ineficiente devido a atrasos na execução de contratos e à duplicação de esforços necessários para manter vários livros. Também é vulnerável porque, se um sistema central (por exemplo, um banco) for comprometido, devido a fraude, ataque cibernético ou um simples erro, toda a rede comercial será afetada

Tais transações vêm aumentando exponencialmente, seja em sistemas bancários, cartões de crédito, transações *e-commerce*, bem como outras transações entre fornecedores e consumidores via *internet*, e proporcionalmente a preocupação com questões relativas a segurança, vulnerabilidade e confiabilidade.

Blockchain é uma tecnologia relativamente nova, também conhecida como *Distributed Ledger Technology* (DTL). Em 2008, um indivíduo com o pseudônimo Satosh Nakamoto (NAKAMOTO, 2008) publicou um *whitepaper* seminal, apresentado ao grupo de discussão "*The Cryptography Mailing*", que delineou o conceito de uma criptomoeda denominada *Bitcoin*, cuja proposta, conforme (FORMIGONE, 2017 p.1), era a criação de uma moeda digital que funcionasse em uma rede peer-to-peer (P2P) que permitisse o envio de pagamentos online de forma totalmente segura, sem o envolvimento de instituições financeiras, para todos os participantes da rede. Nesse sentido, *Blockchain* foi criado motivado pela necessidade de um sistema eficiente, econômico, confiável e seguro para conduzir e registrar transações financeiras. Daí surge a pergunta: qual a relação entre *Blockchain* e *Bitcoin*? *Blockchain* é a plataforma utilizada para o funcionamento da rede *Bitcoin* e de várias outras criptopmoedas. Ainda de acordo com (FORMIGONE, 2017**,** p.3),

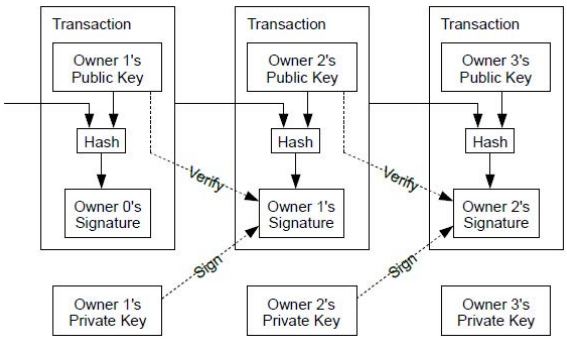
> Após a implantação das primeiras criptomoedas, vários especialistas observaram que propriedades intrínsecas à tecnologia Blockchain (tais como segurança, resiliência, inviolabilidade e imutabilidade) poderiam ser usadas em vários outros tipos de aplicações.

> Neste sentido, as plataformas de desenvolvimento Blockchain evoluíram e permitiram a inserção de transações mais complexas através dos contratos inteligentes denominados (*smart contracts*).

O *Blockchain* deve seu nome à forma como armazena dados de transações - em blocos (*blocks*) que são vinculados para formar uma cadeia (*chain*). *Blockchain*, de acordo com (LEMIEUX, 2017 apud NARAYANAN, 2016, p. 2271),

> é um tipo de tecnologia de contabilidade distribuída na qual, conjuntos confirmados e validados de transações são mantidos em blocos, e os blocos são encadeados de uma forma que dificulta a adulteração e permite que um participante apenas acrescente blocos adicionais de forma sequencial. Assim, cada cadeia começa com um bloco original, ou gênese, seguido por uma sequência ordenada por tempo, de blocos em que, cada bloco contém um *hash* - um número aleatório de 256 bits gerado computacionalmente a partir de informações de entrada do bloco anterior na cadeia. Isso forma uma longa cadeia contínua de *hashes*, daí o nome *Blockchain* (veja a Fig. 1). O comprimento do *Blockchain* é referido como altura do bloco, isto é, a distância entre o gênesis e o último bloco da cadeia.

Figura 01. Estrutura do Blockchain



Fonte: Nakamoto (2008)

Blockchain pode ser visto com uma "*leadger of facts*", ou livro razão, replicada em computadores que participam e uma rede *peer-to-peer* onde, conforme (FORMIGONI, 2017, p. 6):

- O *ledger* é um livro de registros digital, no qual uma vez validado um registro, este nunca mais poderá ser apagado;
- Um fato (*fact*) pode significar várias coisas, desde uma transação monetária, a um conteúdo de determinado documento, ou até mesmo um programa de computador, contendo, em algumas plataformas, até uma base de dados pequena;
- Os membros participantes da rede podem, ou não ser anônimos e são chamados *peers* ou "nós";
- Toda operação ou transação dentro da *ledger* é protegida por tecnologias criptográficas de assinatura digital, inclusive para identificar os nós emissores e receptores das transações;
- Quando um nó deseja adicionar ao *ledger* um fato novo, é necessário um consenso entre todos ou alguns nós previamente determinados da rede, para decidir se um fato pode ser registrado no *ledger*;
- Havendo consenso, o fato será escrito e nunca mais poderá ser apagado, em tese, um processo levemente semelhante à escritura e registro de um imóvel no Brasil.

A tecnologia *Blockchain*, além das criptomoedas, pode ser utilizada em diversas áreas de negócio tais como: serviços financeiros, seguros, saúde, internet das coisas. Como visto, a possibilidade de aplicação da tecnologia Blockchain nas diversas áreas de atividade tem despertado o interesse, não só da área financeira, neste sentido, o renomado jornal *The Economist* (ECONOMIST, 2015) crava a tecnologia Blockchain como A Próxima Grande Coisa, "*The Next Big Thing*". Portanto, devemos acompanhar o desenvolvimento desta tecnologia e identificar de quais formas podemos usufruir dela.

5 CICLO DE VIDA DOS DADOS

Em geral, a vida útil dos dados de pesquisa não se encerra ao final da pesquisa. Tais dados devem ser preservados para compartilhamento e reuso. Neste sentido várias propostas de ciclo de vida de dados científico foram apresentadas com objetivo de descrever todas as etapas do processo de curadoria dos dados.

Dentre os diversos Ciclos de Vida dos Dados Científicos, podemos destacar o ciclo de vida dos dados DataOne (Sayão e Sales, 2015), o Ciclo de Vida de Dados para Ciência da Informação Sant'Ana (Sant'Ana, 2016, p4).

Durante o processo de pesquisa, os dados, muitas vezes, ficam armazenados localmente computadores particulares dos pesquisadores, em algum meio digital de armazenamento ou em repositórios dos laboratórios de pesquisa. Estes dados ficam vulneráveis a perda, uso indevido por pessoas não autorizadas, não permitindo assim que se cumpram todas as fases do ciclo de vida dos dados, conforme o ciclo de vida exemplificado na (Fig. 2). Por outro lado, existe uma necessidade latente dos pesquisadores com relação ao compartilhamento e preservação dos seus dados. Nesta direção, Sales e Cavalcanti (2015) recomendam que sejam adotadas práticas de preservação dos dados digitais e que a curadoria de dados de pesquisa se apresenta como uma solução a ser adotada para se atingir o objetivo de ter segurança na preservação e compartilhamento dos dados, sendo o desenvolvimento de repositórios de dados uma alternativa para tal preservação e o exercício da curadoria digital.

Figura 2. Ciclo de Vida dos Dados

Fonte: Sayão e Sales (2015)

Em 2017, foi publicado no *Digital Science Report* um artigo com título "*Blockchain for Research*", (ROSSUM, 2017), que pretende, conforme dito em sua introdução, ampliar o potencial do *Blockchain* para transformar a comunicação acadêmica e a pesquisa em geral, destacando as possibilidades de como a Tecnologia *Blockchain* pode tocar muitos aspectos críticos da comunicação acadêmica, incluindo transparência, confiança, reprodutibilidade e crédito.

6 POSSÍVEIS USOS DO BLOCKCHAIN PELOS PROFISSIONAIS DA INFORMAÇÃO

A utilização da tecnologia *Blockchain* tem sido objeto de estudos de vários pesquisadores. De acordo com (BRAW, 2018), a tecnologia promete uma maneira nova e descentralizada de gravar e armazenar dados. A aplicação da tecnologia *Blockchain* nas diversas áreas de atividade tem despertado o interesse, não só da área financeira, bem como nas áreas médicas, governo, segurança, dentre outras áreas.

7 *BLOCKCHAIN* E O CICLO DE VIDA DOS DADOS

Devido a sua característica de tornar ativos digitais imutáveis, transparentes, externamente prováveis, validos e distribuídos (e potencialmente permanentes), *Blockchain* apresenta-se como a tecnologia que pode ser utilizada em todo o ciclo de pesquisa, desde sua criação até catalogação e armazenamento, sendo assim, em todas as fases do ciclo de vida podem ocorrer em um sistema Blockchain. Este tópico está baseado em (BARTLING, 2018).

Alguns sistemas de banco de dados possuem características *Blockchain*. BigchainDB é um exemplo de banco de dados distribuído *big data* com características *Blockchain* (controle descentralizado, imutabilidade, e a transferência de ativos digitais). Tais bancos de dados podem ser utilizados para realizar todas as fases do ciclo de vida dos dados além de permitir o acesso aos dados conforme as condições de acesso já apresentadas.

Aquisição de dados

Todos os dados de pesquisa que são adquiridos podem ir para um banco de dados *Blockchain*. Todos os dados que são adquiridos durante um experimento podem estar disponíveis primeiro para um determinado público (*secure access*). Poderia se tornar abertamente disponível e poderia ser reutilizado por outros pesquisadores (*open access*). No entanto, isso não deve necessariamente ser o caso, pois um pesquisador pode controlar quem pode acessar os dados (*managed access*).

Gerenciamento dos dados e análise

A análise dos dados, pós-processamento e estatística pode ser protocolizada no banco de dados *Blockchain* e comprovada aos pares.

Compartilhamento dos dados

Através de bancos de dados *Blockchain*, os dados podem ser armazenados e compartilhados. As tecnologias *Blockchain* podem fornecer uma redundância e disponibilidade de dados, por ex. IPFS (*InterPlanetary File Sys*tem), que de acordo com IPFS (2018), é um novo protocolo de distribuição hipermídia, endereçado por conteúdo e identidades. A criptografia associada pode assegurar que os dados estejam disponíveis apenas para certas pessoas, grupos e a partir de pontos de tempo definidos.

Publicação e arquivamento

A tecnologia *Blockchain* pode ser usada para "assinar" publicações anônimas com credibilidade fornecendo "assinaturas".

Avaliação de pesquisa

Um *Blockchain* (por exemplo, Namecoin) pode ser usado para registrar e manter identificadores de pesquisa exclusivos como ORCID[1] ou *links* para publicações ou conjuntos de dados como o DOI[2].

Fundo de pesquisa

Blockchain poderia permitir muitas novas maneiras de distribuir dinheiro para pesquisa. Por exemplo, o provedor de financiamento de pesquisa pode escolher uma combinação de características de diferentes tipos de comportamento do pesquisador que deseja apoiar.

Blockchain, embora seja uma tecnologia cuja aplicação já é bastante difundida no mercado de criptomoedas, seu uso no ciclo de vida dos dados carece de uma maior investigação.

8 CONSIDERAÇÕES FINAIS

Pelo exposto, percebe-se que a tecnologia *Blockchain*, tem características que se adequam a resolver alguns dos problemas da gestão de dados científicos, cujo problema é a forma multifacetada de sua gestão, onde cada fase do ciclo de vida necessita de uma tecnologia ou método para ser aplicado.

Com a tecnologia *Blockchain*, pode-se envolver todo do ciclo de dados de forma descentralizada desde a criação, gerenciamento, armazenamento, publicação e reuso.

Sendo uma tecnologia nova, muitos desafios ainda deverão ser superados, tais como, a necessidade de um maior conhecimento da tecnologia, adicionar mais recursos, como acesso a banco

1 ORCID - https://orcid.org/ - Open Researcher and Contributor ID
2 DOI - https://www.doi.org/ - Data Object Identifier System

de dados e como medir a produtividade dos pesquisadores, seja através dos institutos de pesquisa ou um banco de dados de pesquisadores.

Olhando para um novo horizonte, o *Blockchain*, como uma tecnologia desruptiva ainda necessita de muitos estudos e investimentos e as universidades, institutos de pesquisa e órgãos fomentadores de pesquisa devem investigar o potencial da tecnologia *Blockchain* e sua aplicação no Ciclo de Vida dos Dados.

REFERÊNCIAS

ALMAN S. e HIRISH S. Ways to Use **Blockchain in Libriries em Blockchain for Information Profession**, 2017 [Online] Disponível em:<https://ischoolblogs.sjsu.edu/Blockchains/Blockchains--applied/applications/>. Acessado em: 09 jul. 2018.

BARTLING S. **Blockchain for Science and Knowledje Creation em Blockchain for Science**. Disponível em:<https://docs.google.com/document/d/1Uhjb4K69l0bSx7UXYUStV_rjuPC7VGo0E-Ra-7xEsr58/edit#>. Acessado em 22 jul. 2018.

BROWN, Jessica Leigth, Blockchain in the Libririan? Researcheres Explore Potential Applications. EdSurge, 2018 [on line]. Disponível em:<https://www.edsurge.com/news/2018-02-01-blockchain--in-the-library-researchers-explore-potential-applications>. Acessado em 09/07/2018.

ECONOMIST The, **The Next Bit Thing: or is it,** 2015. Disponível em: <https://www.economist.com/special-report/2015/05/09/the-next-big-thing>. Acessado em: 01 out 2018.

FORMIGONI, J. R. F.; BRAGA, A. M.; LEAL, R. L. V. **Tecnologia Blockchain: uma visão geral**. 2017. Disponível em: <https://www.cpqd.com.br/wp-content/uploads/2017/03/cpqd-whitepaper--Blockchain-impresso.pdf>. Acessado em: 01 jul. 2018.

GUPTA, M. **Blockchain for Dummies®**, IBM Limited Edition. Wiley Brand John Wiley & Sons, Inc., 2017

IPFS – **The Permanent Web**, 2018. Disponível em:<https://github.com/ipfs/ipfs>. Acessado em 17 set. 2018.

LEMIEUX, V. L. **A Typology Blockchain Recordkeeping Soluctions and Some Reflections on their Implications for the Future of Archival Preservation**, 2017 [online]. Disponível em:<http://dcicblog.umd.edu/wp-content/uploads/sites/13/2017/06/Lemieux.pdf>. Acessado em: 20 jul 2018.

NAKAMOTO, S. **A Peer-to-peer Electronic Cash System**, 2008 Disponível em: <https://bitcoin.org/bitcoin.pdf>. Acessado em: 28 set 2018.

ROSSUM, J. V., **Blockchain for Researh: Perspectives on a New Paradigm for Scholarly Communication**. DIGITAL SCIENCE REPORT, Nov/2017 DOI: https://org/IO.6084/m9.figshare.5607778 Acessado em 02 jul 2018**.**

SALES, L. F., CAVALCANTI, M. T. **Seleção e avalição de coleções de dados digitais de pesquisa: uma possível abordagem metodológica**. Informação & Tecnologia (ITEC): Marília/João Pessoa, 2(2): 88-105, jul./dez., 2015. Disponível em:<http://periodicos.ufpb.br/index.php/itec/article/view/34134/17537>. Acesso em: 29 set. 2018.

SAYÃO, L. F., SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Comissão Nacional de Energia Nuclear: Rio de Janeiro. 2015. Disponível em: http://carpedien.ien.gov.br/bitstream/ien/1624/1/GUIA_DE_DADOS_DE_PESQUISA.pdf. Acesso em: 29 set. 2018.

SOUZA, A. H. R. L.; OLIVEIRA A. F.; D'Avila R. T. **O Modelo de Referência OAIS e a Preservação Digital Distribuída** Ci. Inf., Brasília, DF, v. 41 n. 1, p.65-73, jan./abr., 2012 Acesso em 28 set. 2018

WEB SCRAPING DO RESEARCHERID:
proposta de sistema para o monitoramento de Índice H de pesquisadores no Brasil

WEB SCRAPING IN RESEARCHERID:
proposal to the monitoring system of H-index of the researchers in Brazil

Alexandre Ribas Semeler[1e2], Adilson Luiz Pinto[(2)] , Arthur Oliveira [(3)]
(1) Universidade Federal do Rio Grande do Sul, Instituto de Geociências, Campus do Vale Av. Bento Gonçalves, 9500 - Porto Alegre - RS - Brasil, alexandre.semeler@ufrgs.br.
(2) Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Ciência da Informação, Campus Professor João David Ferreira Lima - Trindade - Florianópolis - Santa Catarina - Brasil - CEP 88.040-900, adilson.pinto@ufsc.br.
(3) Universidade Federal do Rio Grande do Sul, Instituto de Informática, Campus do Vale Av. Bento Gonçalves, 9500 - Porto Alegre - RS - Brasil, arthur.holiver@gmail.com.

Resumo:
A dimensão dos dados aponta uma nova tendência de estudos e práticas que vem sendo adotada por cientistas da informação interessados em estudos métricos que visem o uso de dados de citação e referências. Essas abordagens ressaltam métodos e tecnologias que garantem a interoperabilidade e a criação de uma identidade única para autores e de documentos científicos. O ResearcherID é um sistema de identificação de autores científicos, criado em 2008 pela Thomson Reuters com o objetivo de resolver o problema da identificação de autores. O ResearcherID foi adotado pelo Conselho Nacional de Pesquisa do Brasil como fonte de dados para coleta do Índice H nacional. Este indicador é aplicado para medir a produtividade e para a visualização do impacto de cientistas baseando-se nos seus artigos mais citados. Nesse contexto, a proposta desse estudo será a de medir o índice H dos pesquisadores brasileiros cadastrados no ResearcherID. Os procedimentos metodológicos exigem a aplicação de conhecimentos inerentes a linguagem de programação *Python*. São utilizadas técnicas de web scraping e navegação web automatizada para recuperação de informação no ResearcherID. O resultado do estudo consolida-se na forma de dois *scripts* em Python (Apêndice A e Apêndice B) tais scripts objetivam-se como base para a elaboração de um sistema de monitoramento do índice H no ResearcherID.
Palavras-chave: Dados de Citação. ResearcherID. Índice H. Web Scraping.

Abstract:
The data dimension presents a new trends in studies and practices adopted by information scientists interested in metric studies that aim at the use of data citation and data references. These trends emphasize methods and technologies that ensure interoperability and the creation of a unique identity to authors and scientific documents. The ResearcherID is a system of identification of scientific authors, created in 2008 by Thomson Reuters with the purpose of solving the problem of the identification of authors. The ResearcherID was adopted by the National Research Council of Brazil as data source to validation and collection of the national H-index. This indicator is applied to measure the productivity and visualization of the impact of scientists based on their most cited articles. The aim of this study will be to measure the H-index of Brazilian researchers enrolled in ResearcherID. The methodological procedures require the application of knowledge inherent in the Python programming language. The web scraping and web surfing techniques are used for information retrieval in ResearcherID. The results of the study are consolidated in two Python scripts (Annex A and Annex B). These scripts are intended as a basis for the development of an H-index monitoring system in ResearcherID.
Keywords: Data Citation. ResearcherID. H-index. Web Scraping.

1 INTRODUÇÃO

A dimensão dos dados digitais aponta uma nova tendência de estudos e práticas que vem sendo adotada por cientistas da informação interessados em estudos métricos, em especifico aqueles que visem o uso de dados de citação bibliográfica.

Os estudos métricos com base neste tipo de dados não são algo novo, desde a década de 1963 são desenvolvidos pelo Institute for Scientific Information vide o exemplo do Science Citation Index, uma das bases de dados de citação científica mais importantes do mundo. A identificação precisa de autores e suas citações permite visualizar o seu impacto da produção científica. Isso também permite aos cientistas integrar suas publicações criar uma única identidade científica.

Neste contexto, propomos um estudo que irá utilizar a base de dados de citação do ResearacherID para identificar a variação do Índice H Brasileiro. Existem diversas outras fontes de dados de citação entre outras que fornecem o Índice H, Google Scholar, Scopus, ResearchGate entre outras. No entanto, ressalta-se que o Conselho Nacional de Pesquisas do Brasil utiliza a plataforma do ResearcherID para visualizar o número de citações e o Índice H dos pesquisadores Brasileiros.

O (CNPq) adotou o ResearcherID para validar os dados sobre citações a publicações científicas de impacto internacional na sua plataforma de currículos Lattes, vide Apêndice F. O CNPq utiliza o ResearcherID, para coletar e validar automaticamente os dados sobre o número total de citações, número de trabalhos e Índice H, dos pesquisadores que possuem publicações científica de impacto internacional na Web of Science por meio do ResearcherID.

Segundo Hirsch (2005), o Índice H é um indicador de quantificação para a produtividade e visualização do impacto de cientistas baseando-se nos seus artigos mais citados.

O resultado de busca por informação em dados de citação, principalmente, quando o foco está Índice H, pode indicar uma frente de pesquisa. Uma frente de pesquisa é composta pelo conjunto de autores mais citados em um determinado campo e indica a elite de autores de um campo científico e ou um pais. Nesse contexto, os dados de citação podem refletir a produtividade e a visibilidade dos autores em diversos campos científicos.

Tendo em vista o desenvolvimento das bases de dados de citação, conforme explanado por Rice e Southall (2016), a Clarivate Analitics desenvolveu em 2008 ResearcherID sistema utilizado para rastrear citações em conjuntos de dados da literatura científica de impacto internacional.

O ResearcherID é um sistema de identificação de autores científicos, criado pela Thompson Reuters com o objetivo de resolver o problema da identificação de autores. O ResearcherID se integra com a Web of Science (Science Citation Index), é um identificador digital único, gratuito e persistente, que distingue um pesquisador de outro e resolve o problema da ambiguidade e semelhança de nomes de autores e indivíduos, substituindo as variações de nome por um único código numérico facilitando o registro de informações ele serve para automatizar a atualização das publicações e de seus dados de citação possibilitando a interoperabilidade entre as publicações e suas citações a partir de uma única conta. (RESEARCHERID, 2018)

Esse ID científico é formado por um código de 11 dígitos, ex.: (J-9183-2016), na forma de caracteres alfanuméricos, sendo único para cada pesquisador cadastrado. O ResearcherID impede ambiguidades na identificação de autores e colaboradores em publicações. Seu funcionamento é

semelhante ao do Digital Object Identifier (DOI), código para objetos como artigos científicos, teses e dissertações (RESEARCHERID, 2018; RICE; SOUTHALL, 2016).

O ResearcherID é um identificador único que permite que os pesquisadores gerenciem suas listas de publicações, rastreiem suas contagens de citações e Índice h. (RESEARCHERID; 2018)

A utilização desses IDs favorece a localização e a contagem das publicações, além de reunir os dados de citações recebidas pelo conjunto de trabalhos de sua autoria possibilitando a visualização do Índice H.

Nesse sentido, a proposta desse estudo será a de medir o Índice H dos pesquisadores brasileiros cadastrados no ResearcherID, contar o número de publicações e visualizar o número de publicações com citação desses pesquisadores.

A seguir apresentam-se os objetivos, os procedimentos metodológicos e os resultados desse estudo com maiores detalhes.

2 OBJETIVOS

O objetivo geral desta proposta é desenvolver scripts em Python para monitorar o Índice H de pesquisadores brasileiros no ResearcherID. Para isso, propõem-se os seguintes objetivos específicos:

a) automatizar a coleta do Índice H de pesquisadores registrados no ResearcherID;
b) criar *scripts* de web scraping e navegação automatizada em linguagem de programação Python para coleta de dados de citação;
c) visualizar o Índice H dos pesquisadores brasileiros cadastrados no ResearcherID, contar o número de publicações e visualizar o número de suas publicações com citação.

3 PROCEDIMENTOS METODOLÓGICOS

O principal método adotado para este estudo foi o *web scraping*, método que se define como processo de extração e combinação de conteúdos web de forma sistemática e automatizada. Em tal processo, um agente de *software*, o *scraping*, simula o comportamento de navegação humano em servidores web copiando e reorganizando dados desorganizados em dados organizados (GLEZ-PENA et al., 2013).

Existem diferentes técnicas de web *scraping:* a manipulação de HTTP, a mineração de dados, as ferramentas *scraping*, a cópia manual e os microformatos. A manipulação de HTTP permite a colheita de dados estáticos e dinâmicos de um *site* através de uma solicitação HTTP. A mineração de dados é um processo automático que reconhece as informações de um site de acordo com *scripts* predefinidos que contêm dados incorporados. Ferramentas de *scraping* são *softwares* utilizados para extrair informações relacionadas a sites ou a funcionalidades e/ou estruturas de dados na *web*. As ferramentas de *scraping* também servem para extrair dados de redes sociais e são úteis para todos os tipos de atividades de *marketing* web. A cópia manual, embora existam ferramentas de *scraping* disponíveis, às vezes é necessária, por exemplo, em alguns casos em que a informação do site for bloqueada contra qualquer

forma web *scraping*. Por fim, os microformatos são referentes às tecnologias da web semântica, como dados abertos vinculados via RDF. Esses são conjuntos de informações geralmente elaboradas para serem intercambiadas via vocabulários e ontologias (WEBSTER, 2015).

O funcionamento de um web *scraping* ocorre da seguinte maneira: (a) acesso ao site via protocolo de comunicação web (HTTP), durante o processo de solicitação e resposta entre um cliente, normalmente um navegador da web, e um servidor da web; (b) o programa de *scraper* analisa o HTML e pode extrair seu conteúdo; (c) armazena o *output* em conteúdo textual, com o objetivo de transformar o conteúdo extraído em uma representação estruturada para posterior análise e armazenamento (GLEZ-PENA et al., 2013).

O objetivo da web *scraping* é transformar os dados desestruturados da *web* em representações estruturadas, como aquelas de formatos tabulares, por exemplo: os tipos de arquivos *Comma Separated Values* (CSV), *Tab Separated Values* (TSV), e/ou e*Xtensible Markup Languag*e (XML). Esses formatos permitem a analise do conteúdo de sites da mesma forma que permitem salvar a navegação feita pelos browsers web, como o Chrome ou FireFox.

No que segue apresenta-se uma sumarização dos procedimentos metodológicos, conforme o Quadro 1.

Quadro 1 – Caracterização do método

CARACTERIZAÇÃO DO MÉTODO	
TIPO DE PESQUISA	Exploratória e Descritiva
ESTRATÉGIA	Uso de linguagens de programação para coletar dados de citação
NATUREZA	Quantitativa
PANORAMA GERAL DA INVESTIGAÇÃO	Índice H de pesquisadores brasileiros
AMOSTRA	43.114 - IDS em 17 set. 2018
CORPUS TOTAL	43.114 - IDS em 17 set. 2018
INSTRUMENTOS DE COLETA	*Web scraping* e navegação web automatizada
SOFTWARES	*IDE* = *Pycharm* *Python*=2.7 *Módulos* = Codecs, BeautifulSoup, Selenium
FONTE DE COLETA DE DADOS	ResearcherID

Fonte: Dados da Pesquisa (2018).

Com base nesses procedimentos apresentam-se os resultados da pesquisa.

4 RESULTADOS

Os resultados obtidos nesse estudo representam um possível modelo para execução automatizada do processo de coleta de dados de citação oferecidos pelo ResearcherID. Os resultados subdividem-se em um plano técnico, criação de *scripts* para web *scraping* e navegação web automática

das informações no ResearcherID e em um plano metodológico de descrição e interpretação dos dados coletados sobre o Índice H brasileiro.

4.1 Resultados técnicos

Como resultado técnico obteve-se dois *scripts* em Python 2.7.15 os quais são aplicados para o monitoramento do Índice H dos pesquisadores brasileiros cadastrados no ResearcherID, contar o número de publicações e visualizar o número de publicações com citação desses pesquisadores com base no núcleo da Web of Science. Os mesmos podem ser visualizados no apêndice A e no apêndice B deste trabalho.

O apêndice A, apresenta o *script1*, o qual executa a primeira coleta de dados no ResearcheriD. O *script1* automatiza o web *scraping* das informações sobre: o nome de pesquisadores, sua instituição, o código ID e as áreas do conhecimento relacionadas ao seu ID. O *script1* salva uma lista contendo, o um conjunto de dados com todos os IDS de pesquisadores brasileiros que possuem registro no ResearcherID, a lista contém (43.114) pesquisadores brasileiros registrados em 19 de novembro de 2018. O script1 possui um dispositivo de tratamento de falhas, que garante que quando interrompida a extração, em caso de *timeout* na página, a extração seja reiniciada no último ponto valido coletado no Researcher. O script1 também evita as duplicações de IDs. O script1 gera um arquivo de log que contém apenas o último índice a ser extraído da lista usada. No começo da execução o script1 testa se já existe um arquivo com um índice, se o arquivo estiver vazio/não existir o índice a ser usado será 1 por padrão. Ao final da execução, caso o programa não seja interrompido por erro ou parada intencional do usuário, haverá no log de execução a data e hora em que a extração foi finalizada.

Os resultados da coleta feita pelo script1 são salvos em um arquivo *.TSV* . Está coleta e contém as seguintes informações: Nome do pesquisador, a instituição, o código no ResearcherID e as palavras chave. Estes dados servem para o *script2* como base para uma nova coleta de dados sobre as métricas de citação do *ResearcherID*.

O apêndice B representa o *script2,* o qual lê a lista de dados escrita pelo *script1* e executa a coleta de informações sobre as métricas de citação do ResearcherID. *O script2* lê os registros do arquivo .TSV e gera links *(http://www.researcherid.com/rid/+ ID* do pesquisador) usando os IDs do arquivo .TSV. Estes Links são únicos e representam os dados de citação de cada pesquisador. O *script2* possui o mesmo mecanismo de falhas do *script1*. O script2 é paralelizado em 8 processos iguais, cada processo coleta informações de (1/8) é feito de maneira aleatória na lista de links gerados a partir do do arquivo .TSV gerado pelo *script1*. As informações coletadas pelo script2 snao salvas em 8 arquivos .TSV diferentes durante a extração para acelerar o processo de coleta, no final da coleta as informações desses 8 arquivos são salvas em um único arquivo .TSV contendo os dados de citação da base, em especial: ID, total de publicações na lista, publicações com dados de citação, a soma de citações no tempo, a média de citações por artigo e o Índice H dos pesquisadores brasileiros. Por fim, o *script2* soma as publicações com dados de citação e faz a média do Índice H.

4.2 Interpretações da coleta

Os resultados da coleta permitem não apenas a visualização do Índice H, mas também outras informações apresentadas pelo ResearcherID como: total de publicações na lista, publicações com dados de citação, a soma de citações no tempo, a média de citações por artigo e o Índice H.

A descrição e a intepretação dos dados coletados permitiu visualizar Índices H discrepantes e aqueles que não correspondem a autores únicos, como aqueles utilizados para medir instituições ou mesmo aqueles feitos erroneamente. Estes dados foram excluídos da analise por não representarem o objetivo deste trabalho.

Assim, a coleta permitiu visualizar algumas situações, conforme o apêndice C, de um total de (40.114) pesquisadores cadastrados no ResearcherID, (9560) pesquisadores não possuem métricas no ResearcherID (N/A), ou seja nunca foram citados no Science Citation Index. Outros (960) possuem publicações com zero citação, ou seja, estes trabalhos não possuem impacto na WOS ou ainda não foram citados. As (10) maiores frequências de índice a H estão entre 1 e 10: (1=4006, 2=3454,3=3244, 4=2752, 5=2572, 6=2141, 7=1823, 8=1651, 9=1410, 10=1231)

Com base nos (40.114) IDs a média do Índice H nacional é (7,97), levando-se em consideração pesquisadores únicos e aqueles que possuem índice H maior que zero, ou seja, um montante de (32.255) pesquisadores brasileiros.

A frente de pesquisa nacional pode ser visualizada no apêndice D e concentra-se no campo cientifico da Física**.** Conforme a Tabela 2 contida no apêndice C, o ID (B-2946-2012) possui o maior índice H nacional válido (96). Este índice é de um professor do Instituto de Física Gleb Wataghin da Universidade Estadual de Campinas (UNICAMP). O segundo maior índice H (92) também corresponde a área de Física, conforme o ID (L-6239-2016), pertence a um professor do Instituto de Física da Universidade de São Paulo. Ambos atuam na área de Física Nuclear e são Bolsistas de Produtividade em Pesquisa do CNPq - Nível 1D.

O apêndice D também permite visualizar e ranquear os pesquisadores brasileiros conforme a variação dos Índices H mais altos (mínimo 75 e máximo 96) são únicos. O ID que recebeu mais citações (39544) é (L-1621-2016) de um professor da Universidade Estadual do Rio de Janeiro que também atua na área de Física nuclear sendo Bolsista de Produtividade em Pesquisa do CNPq - Nível 1C.

Já as publicações com citação na WOS compreendem o montante de (877.624) estas são publicações que receberam no mínimo 1 citação em periódico com fator de impacto. O número total de publicações registradas nas listas foi (1.146.855), no entanto este número é prejudicado ao representar o número de produções do Brasil na WOS, pois alguns pesquisadores não preencheram seus IDs apenas com a sua produção na WOS mas com outras fontes de dados de citação como Google Scholar e Scopus.

Por fim, o ID (K-5168-2013) possui a maior média de citações recebidas (99.88) no ReseacherID a coordenadora do Núcleo de Pesquisa em Asma e Inflamação das Vias Aéreas (NUPAIVA), com índice H (22) a pesquisadora da Universidade Federal de Santa Catarina valida toda sua produção na Plataforma Lattes.

No entanto, chamou a atenção que os maiores índices H brasileiros (140), (105) e (103) não serem válidos, pois foram realizados com objetivos diversos e não com o objetivo de validar os

dados de citação no Lattes CNPq. É necessário ter em mente que o ReserchерID é um identificador pessoal de autor e possui o objetivo de fornecer métricas para produção de autores únicos e não de instituições, ressaltando também que não é uma ferramenta para pesquisa bibliográfica.

Assim o índice H 140, (A-9780-2017), atribuído ao Instituto de Física, Gleb Wataghin da UNICAMP e o índice H 105, (M-2664-2016), que representa o Instituto de Física de São Carlos da Universidade de São Paulo prejudicam a visualização do índice H nacional de autores já que são índices de um conjunto de autores de uma instituição. Outra discrepância é o índice H 103 (E-4724-2015), de um estudante do Laboratório de Raioisótopos do Centro de Energia Nuclear na Agricultura – CENA, que foi construído de maneira errônea pois o autor em sua lista pessoal de publicações insere produções científicas de outros autores de certo ele utilizou a plataforma para visualizar a produção em sua área de pesquisa desconhecendo o objetivo do ResearcherID, tão pouco ele valida seu Lattes com os dados.

Estes três casos ressaltam a relevância desta proposta de trabalho. Pois a visualização do Índice H nacional é usada pelo CNPq como estratégia para fomentar e desenvolver novas pesquisas, pautando-se na relevância e no impacto dos pesquisadores atribuída pela WOS, nas mais diversas áreas do conhecimento. No entanto, é necessário fazer o uso correto da plataforma ResearcherID que é um instrumento de validação de citações utilizada pela plataforma Lattes do CNPq.

4 CONSIDERAÇÕES FINAIS

O Índice H é usado pelo CNPq como um dos indicadores para pautar a distribuição de Bolsas de Pesquisa e para pautar a relevância acadêmica da produção cientifica no Brasil. Conhecer o Índice H nacional torna-se tarefa de suma relevância. Pois o Índice H indica o impacto e a relevância dos trabalhos mais citados ao longo do tempo. Revelando a visibilidade e o uso da produção científica de impacto internacional de cada pesquisador.

Considera-se esse estudo como uma estratégia fundamental para medir a variação do Índice H nacional, a variação da média de citações de impacto internacional, medir a soma das citações da produção científica nacional, assim como para visualizar o total de publicações nacionais nas listas do ResearcherID com citações.

Por fim, ressalta-se que os dados de citação e os IDs de autores são ferramentas necessária para visualização do impacto internacional do que é produzido em nosso país, sendo automatização desse processo de suma aplicabilidade em estudos métricos que visem o uso de dados para qualificar a produção cientifica nacional. Conclui-se que os usuários brasileiros deveriam ter maior critério ao utilizar o ResearcherId, pois a ferramenta é utilizada pelo CNPq para validar o impacto da produção nacional na Web of Science uma das principais e mais respeitadas bases de dados e documentos bibliográficos.

Agradecimentos

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

REFERÊNCIAS

HEIRICH, J. An index to quantify an individual's scientific research output. **PNAS, v.** 102, n. (46), 2005. Disponível em: https://doi.org/10.1073/pnas.0507655102>. Acesso em: set. 2018.

GLEZ-PEÑA, D. et al.. Web scraping technologies in an API world. **Briefings in Bioinformatics**, v. 15, n. 5, p. 788-797, 2013. Disponível em: <http://bib.oxfordjournals.org/content/15/5/788>. Acesso em: set. 2018.

ResearcherID. Disponível em: < http://www.researcherid.com > Acesso em: set. 2018.

RICE, R.; SOUTHALL, S. **The data librarian's handbook**. London: Facet Publishing, 2016.

WEBSTER, S. **What Is Scraping?** The Basics For Everyone. 2015. Disponível em: https://myhelpster.com/what-is-scraping-the-basics-for-everyone/. Acesso em: set. 2018.

PYTHON. Disponível em: <https://www.python.org/>. Acesso em set. 2018.

Apêndice A – ***Script1:*** coleta de informações gerais no ResearcherID

```
# coding: utf-8
"""
Created on Sep 17 2018
"""

import codecs
from bs4 import BeautifulSoup
import selenium.common.exceptions
# from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as ec
from selenium.webdriver import Firefox
# import os

url = 'http://www.researcherid.com/ViewProfileSearch.action'    # define url alvo
# Caso a execução seja feita com Chrome
# chromedriver = "../scripts/chromedriver"            # path do chromedriver
# os.environ["webdriver.chrome.driver"] = chromedriver
# driver = webdriver.Chrome(chromedriver)                         # abre browser

# abre arquivo para escrita dos resultados
with codecs.open('../dados/csv/RESEARCHID_Extraction2.tsv', 'w', 'utf-8-sig') as re-
sults:
    results.write('Name\tInstitution\tResearcher ID\tKeywords\n')   # escreve headers
results.close()
id_list = []
counter = 0     # inicializa contador de resultados e erros
errors = 0

driver = Firefox()
while True:
    try:
        driver.get(url)                                                 # navega para a
url
        # espera carregamento da combobox para escolher o Pais alvo
        WebDriverWait(driver, 10).until(ec.presence_of_element_located((By.CSS_SELEC-
TOR,
                                                                        '#country >
option:nth-child(33)'))).click()
        break
    except selenium.common.exceptions.TimeoutException:
        continue

# clica para submeter pesquisa
driver.find_element_by_css_selector('#submitImage > img').click()
# define limite de paginas para serem acessadas nos resultados
page_limit = int(WebDriverWait(driver, 10).until(ec.presence_of_element_located(
    (By.CSS_SELECTOR, '#criteria_requestedPage_total'))).text) // 50 + 1
# espera combobox carregar para mostrar 50 resultados por pagina
WebDriverWait(driver, 10).until(ec.presence_of_element_located((By.CSS_SELECTOR,
                                                               '#resultsPerPage >
option:nth-child(3)'))).click()

for page in range(page_limit):      # percorre paginas de resultados da busca
    # espera tabela de resultados ser carregada
    WebDriverWait(driver, 10).until(ec.presence_of_element_located(
        (By.CSS_SELECTOR, '#resultContainer > table:nth-child(23) > tbody > tr:nth-
-child(2) > td > table > tbody')))
    # define raiz da arvore de resultados para parsing
    root = BeautifulSoup(driver.page_source, 'lxml').find('td', {'class': 'resultsTa-
bleBox'}).findNext('tbody')
    with codecs.open('../dados/csv/RESEARCHID_Extraction2.tsv', 'a', 'utf-8-sig') as
results:
        for item in root.findAll('tr')[1:]:     # percorre itens na tabela
            data = item.findAll('td')           # lista todos os resultados
            rid = data[4].getText().replace('\n', '').replace(' ', '')
            if rid in id_list:
                continue
            else:
                id_list.append(rid)
                # escreve resultados no arquivo de output
```

```
                results.write('%s\t%s\t%s\t%s\n' % (data[1].a.getText().replace('\n',
'),
                                                    data[2].getText().replace('\n',
'),
                                                    rid,
                                                    data[5].getText().replace('\n',
')))
                counter += 1        # incrementa contador de resultados
        results.close()
    if page < page_limit:       # testa se as paginas de resultados ja acabaram
        # clica em next page caso contrario
        WebDriverWait(driver, 10).until(ec.presence_of_element_located(
            (By.CSS_SELECTOR, '#resultContainer > table:nth-child(22) > tbody > tr >
td:nth-child(2) > table > '
                              'tbody > tr > td:nth-child(7) > span > img'))).click()
    print counter
    try:        # tenta esperar pelo carregamento de um item da pagina como teste
        WebDriverWait(driver, 20).until(ec.presence_of_element_located(
            (By.ID, 'pageNumnull')))
    except selenium.common.exceptions.TimeoutException:     # em caso de timeout,
trata o erro
        driver.get(url)     # navega para a url alvo novamente
        errors += 1         # incrementa counter de erros
        # reinicia a pesquisa
        WebDriverWait(driver, 30).until(ec.presence_of_element_located(
            (By.CSS_SELECTOR, '#country > option:nth-child(33)'))).click()
        driver.find_element_by_css_selector('#submitImage > img').click()
        WebDriverWait(driver, 40).until(ec.presence_of_element_located(
            (By.CSS_SELECTOR, '#resultsPerPage > option:nth-child(3)'))).click()
        page -= 1       # reduz contador de paginas navegadas em 1
        element = WebDriverWait(driver, 10).until(ec.presence_of_element_located((By.
XPATH, '//*[@id="pageNumnull"]')))
        element.clear()     # limpa caixa de texto com numero de paginas
        element.send_keys('%s' % page)      # escreve numero de paginas na caixa de
texto
        driver.find_element_by_xpath('/html/body/div[3]/form[2]/table/tbody/tr/td/div/
table/tbody/tr[2]/td[2]/div/div/'
                                     'table[1]/tbody/tr/td[2]/table/tbody/tr/td[5]/a/
img').click()  # clica em "go"
driver.quit()
print errors, 'errors happened. Extraction finished.'
```

Fonte: Dados da Pesquisa (2018).

Apêndice B – Script2: coleta das métricas de citação do ResearcherID

```
# coding: utf-8
"""
Created on Aug 27 2018
"""
import codecs
from bs4 import BeautifulSoup
from multiprocessing import Pool
import selenium.common.exceptions
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as ec
from selenium.webdriver.firefox.options import Options
from selenium.webdriver import Firefox
# import os
# from selenium.webdriver.chrome.options import Options
# from selenium import webdriver


################################################################################################
###############
master = codecs.open('../dados/csv/RESEARCHERID-ALL-METRICS.tsv', 'r', 'utf-8')
id_list = master.readlines()[1:]
master.close()
for x in range(len(id_list)):
    id_list[x] = '%s' % id_list[x].split('\t')[0][1:]
################################################################################################
###############
options = Options()
options.add_argument("--headless")


class Control:
    def __init__(self):
        self.counter = 0
        self.article_accumulator = 0
        self.with_citation_accumulator = 0
        self.h_index_accumulator = 0
        self.existent_h_index_counter = 0


def id_cleaning(a_link):
    if '\r\n' in a_link:  # testa finais de linha para verificar se usam \n ou \r\n
        return a_link.replace(' ', '').replace('\r\n', '\t')  # troca final por \t e tira
blank spaces
    else:
        return a_link.replace(' ', '').replace('\n', '\t')


def get_metrics_info(a_soup):   # function que extrai dados alvo da pagina
    return (a_soup.find(id='metrics_totalArticleCount').getText(),
            a_soup.find(id='metrics_articleCountForMetrics').getText(),
            a_soup.find(id='metrics_timesCited').getText(),
            a_soup.find(id='metrics_averagePerItem').getText(),
            a_soup.find(id='metrics_hindex').getText())


def parallel_scraping(a_trio):
    global id_list
    driver = Firefox(firefox_options=options)
    # driver = webdriver.Chrome(chromedriver)
    for item in a_trio[0]:  # for para percorrer a lista de links
        rid = id_cleaning(item.split('\t')[2])
        if rid in id_list:
            continue
        link, name = 'http://www.researcherid.com/rid/' + rid, item.split('\t')[0]
        a_trio[2].counter += 1  # incrementa counter para acompanhar progresso
        print a_trio[2].counter, ': Extracting: %s' % link

        driver.get(link)  # acessa link alvo
        try:
```

```
            driver.find_elements_by_xpath('/html/body/div[3]/div/table/tbody/tr/td[1]/tab-
le/tbody/tr[1]/td[2]/'
                                          'div[1]/ul[1]/''li[2]/a')[0].click()
            # clica para mostrar citation metrics
        except IndexError:
            print '%s apparently got no citation metrics' % link  # exception tratada
            continue
        a_file = codecs.open(a_trio[1], 'a', 'utf-8-sig')
        try:
            WebDriverWait(driver, 10).until(ec.presence_of_element_located(
                (By.CSS_SELECTOR, '#metrics_hindex')))
        except selenium.common.exceptions.TimeoutException:
            a_file.write(link.split('/')[-1] + '\t' + 'TO\tTO\tTO\tTO\tTO\n')
        else:
            root = BeautifulSoup(driver.page_source, 'lxml')  # parsing inicial
            data = get_metrics_info(root)
            if data != '':
                a_trio[2].article_accumulator += int(data[0])
                a_file.write(link.split('/')[-1] + '\t' + '%s\t%s\t%s\t%s\t%s\n' % data)
                if data[-1] > 0:
                    a_trio[2].with_citation_accumulator += int(data[1])
                    a_trio[2].h_index_accumulator += float(data[-1])
                    a_trio[2].existent_h_index_counter += 1
            else:
                a_file.write(link.split('/')[-1] + '\t' + 'NA\tNA\tNA\tNA\tNA\n')
                # aguarda carregamento da pagina para escrever no arquivo de output
        finally:
            a_file.close()
    driver.quit()  # fecha browser


if __name__ == '__main__':
    pool = Pool(processes=8)
    # geckodriver = "../scripts/geckodriver"    # path do geckodriver para firefox

    # chrome_options = Options()
    # chrome_options.add_argument("--headless")
    # chromedriver = "../scripts/chromedriver"    # path do chromedriver
    # os.environ["webdriver.chrome.driver"] = chromedriver

    with codecs.open('../dados/csv/RESEARCHID_Extraction2.tsv', 'r', 'utf-8-sig') as re-
search:      # abre arquivo
        link_list = research.readlines()[1:]    # extrai linhas do arquivo em uma lista
    research.close()                            # fecha arquivo

    with open('../dados/error_logs/ResearcherIDanonScraping.txt', 'w') as log:
        out_file1 = codecs.open('../dados/csv/RESEARCHIDresultsPart1-8.tsv', 'w', 'utf-
-8-sig')  # abre arquivo de output
        out_file2 = codecs.open('../dados/csv/RESEARCHIDresultsPart2-8.tsv', 'w', 'utf-
-8-sig')  # abre arquivo de output
        out_file3 = codecs.open('../dados/csv/RESEARCHIDresultsPart3-8.tsv', 'w', 'utf-
-8-sig')  # abre arquivo de output
        out_file4 = codecs.open('../dados/csv/RESEARCHIDresultsPart4-8.tsv', 'w', 'utf-
-8-sig')  # abre arquivo de output
        out_file5 = codecs.open('../dados/csv/RESEARCHIDresultsPart5-8.tsv', 'w', 'utf-
-8-sig')  # abre arquivo de output
        out_file6 = codecs.open('../dados/csv/RESEARCHIDresultsPart6-8.tsv', 'w', 'utf-
-8-sig')  # abre arquivo de output
        out_file7 = codecs.open('../dados/csv/RESEARCHIDresultsPart7-8.tsv', 'w', 'utf-
-8-sig')  # abre arquivo de output
        out_file8 = codecs.open('../dados/csv/RESEARCHIDresultsPart8-8.tsv', 'w', 'utf-
-8-sig')  # abre arquivo de output
        file_list = [out_file1, out_file2, out_file3, out_file4, out_file5, out_file6, out_file7,
out_file8]
        for out_file in file_list:
            out_file.write('ResearcherID\tTotal in Pub. List\tWith Citation Data\tSum of
Times Cited\t'
                           'Average Citations/Article\tH-index\n')   # escreve headers
            out_file.close()
        control1, control2, control3, control4, control5, control6, control7, control8 =
Control(), Control(), \

Control(), Control(), \
```

```
    Control(), Control(), \

    Control(), Control()
            control_list = [control1, control2, control3, control4, control5, control6, con-
    trol7, control8]
            print 'Begin of extraction,', len(link_list) - 1, 'items to go'

            octan = len(link_list) // 8
            pool.map(parallel_scraping, [(link_list[:octan-1], '../dados/csv/RESEARCHIDresul-
    tsPart1-8.tsv', control1),
                                         (link_list[octan:2*octan-1], '../dados/csv/RESEARCHI-
    DresultsPart2-8.tsv', control2),
                                         (link_list[2*octan:3*octan-1], '../dados/csv/RESEAR-
    CHIDresultsPart3-8.tsv', control3),
                                         (link_list[3*octan:4*octan-1], '../dados/csv/RESEAR-
    CHIDresultsPart4-8.tsv', control4),
                                         (link_list[4*octan:5*octan-1], '../dados/csv/RESEAR-
    CHIDresultsPart5-8.tsv', control5),
                                         (link_list[5*octan:6*octan-1], '../dados/csv/RESEAR-
    CHIDresultsPart6-8.tsv', control6),
                                         (link_list[6*octan:7*octan-1], '../dados/csv/RESEAR-
    CHIDresultsPart7-8.tsv', control7),
                                         (link_list[7*octan:], '../dados/csv/RESEARCHIDresul-
    tsPart8-8.tsv', control8)])
        total_H_index = 0
        total_existent_H_index = 0
        total_cited = 0
        total_articles = 0
        elements_counter = 0
        for x in control_list:
            total_H_index += x.h_index_accumulator
            total_existent_H_index += x.existent_h_index_counter
            total_cited += x.with_citation_accumulator
            total_articles += x.article_accumulator
            elements_counter += x.counter

        log.write('Total count: %s\n' % str(elements_counter))
        log.write('Average H-Index: %s\n' % str(total_H_index / total_existent_H_index))
        log.write('Total articles with citation data: %s\n' % str(total_cited))
        log.write('Total articles: %s\n' % str(total_articles))
        log.close()
        print 'Extraction finished:', elements_counter, 'elements listed'
```

Fonte: Dados da Pesquisa (2018).

Apêndice C – Tabela 1 – Rank de Frequências do Índice H Brasileiro

H index	Freq.	H index	Freq.
0	960	41	17
1	4006	42	25
2	3454	43	5
3	3244	44	21
4	2752	45	5
5	2572	46	12
6	2141	47	10
7	1823	48	12
8	1651	49	16
9	1410	50	7
10	1231	51	6
11	1082	52	3
12	931	53	9
13	748	54	9
14	671	55	7
15	604	56	7
16	526	57	1
17	498	58	7
18	369	59	5
19	355	60	4
20	307	61	3
21	253	62	6
22	251	63	1
23	205	67	3
24	183	68	2
25	159	69	5
26	126	71	2
27	143	73	1
28	96	75	3
29	85	76	3
30	90	77	2
31	66	80	2
32	46	83	2
33	57	84	2
34	37	85	1
35	42	86	1
36	38	92	1
37	22	96	1
38	26	103	1
39	15	105	1
40	28	140	1
		N/A	9576

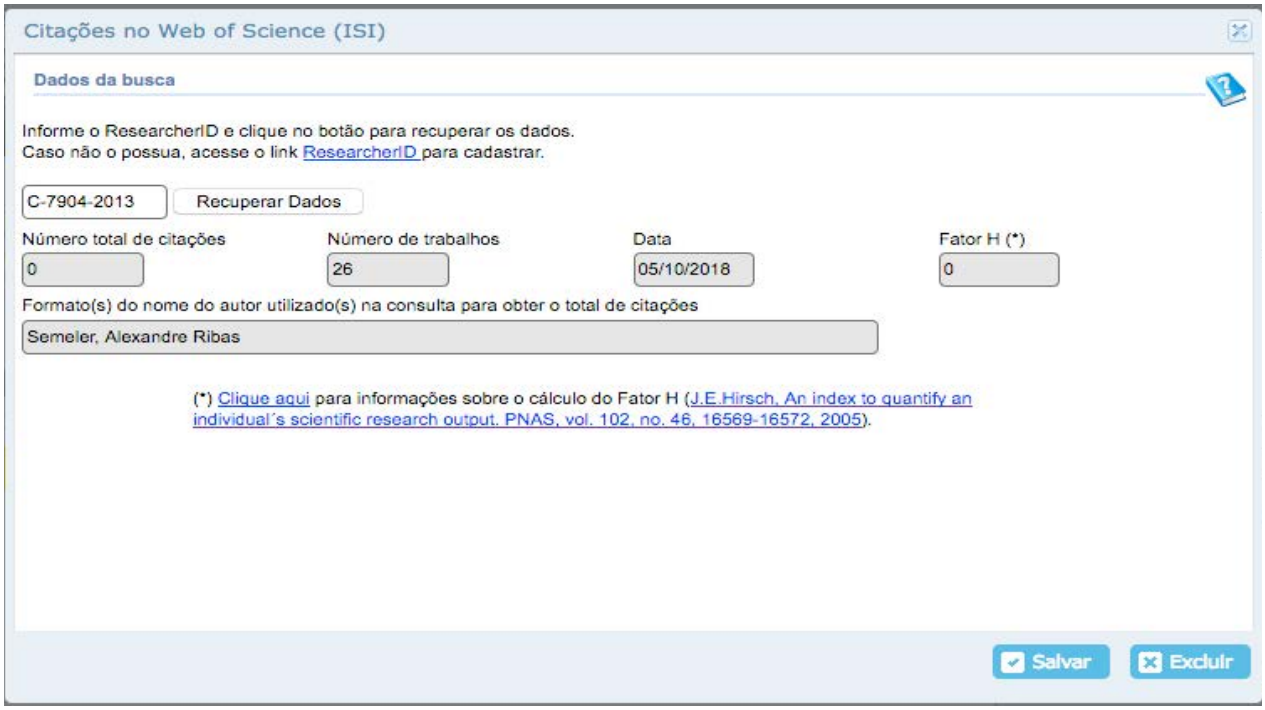
Fonte: Dados da Pesquisa (2018).

Apêndice D – Frente de Pesquisa Nacional Conforme Índice H

Rank/ ID		Total in Pub. List	With Citation Data	Sum of Times Cited	Average Citations/ Article	H-index	Área
1	B-2946-2012	438	438	31.845	72.87	96	Física (UNICAMP)
2	L-6239-2016	408	406	29.667	73.07	92	Física (USP)
3	D-3532-2012	833	820	39.622	48.50	86	Física (UNESP)
4	L-1621-2016	866	745	39.544	53.22	85	Física (UERJ)
5	C-4007-2013	619	619	37.172	60.15	84	Física (UFBA)
6	D-4476-2013	520	520	27.898	53.65	84	Medicina (UFPEL)
7	C-7679-2016	611	298	34.566	115.99	83	Medicina (UFRJ)
8	G-9573-2012	755	755	25.892	34.29	80	Psicologia (UFRGS)
9	L-4142-2016	501	488	28.821	59.30	77	Física (UERJ)
10	E-8874-2010	459	321	18.207	56.72	75	Física (USP)

Fonte: Dados da Pesquisa (2018).

Apêndice F – Validação do Índice H no Lattes CNPq



Fonte: Dados da Pesquisa (2018).