UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

JOYCE MIRELLA DOS ANJOS VIANA

REPRESENTAÇÃO COLABORATIVA DE DADOS CIENTÍFICOS:

Estudo na Rede de Repositórios de Dados Científicos
do Estado de São Paulo

São Carlos - SP
2020

JOYCE MIRELLA DOS ANJOS VIANA

REPRESENTAÇÃO COLABORATIVA DE DADOS CIENTÍFICOS:

Estudo na Rede de Repositórios de Dados Científicos
do Estado de São Paulo

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de São Carlos como requisito parcial para obtenção do título de Mestre em Ciência da Informação.

Área de concentração: Conhecimento, Tecnologia e Inovação.

Linha de pesquisa: Tecnologia, Informação e Representação.

Orientadora: Profa. Dra. Paula Regina Dal'Evedove

Financiamento: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)

São Carlos - SP
2020

[VERSO DA FOLHA DE ROSTO]

VIANA, Joyce Mirella dos Anjos

Representação Colaborativa de Dados Científicos: Estudo na Rede de Repositórios de Dados Científicos do Estado de São Paulo / Joyce Mirella dos Anjos Viana. -- 2020.
127 f. : 30 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador: Paula Regina Dal'Evedove
Banca examinadora: Guilherme Ataíde Dias, Luciana de Sousa Gracioso
Bibliografia

1. Representação colaborativa da informação. 2. Big Data. 3. Dados científicos. 4. Repositórios de dados. 5. Rede de Repositórios de dados científicos do estado de São Paulo. I. Orientador. II. Universidade Federal de São Carlos. III. Título.

Ficha catalográfica elaborada pelo Programa de Geração Automática da Secretaria Geral de Informática (SIn).
DADOS FORNECIDOS PELO(A) AUTOR(A)
Bibliotecário(a) Responsável: Ronildo Santos Prado – CRB/8 7325

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós Graduação em Ciência da Informação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Joyce Mirella dos Anjos Viana, realizada em 06/05/2020:

Profa. Dra. Paula Regina Dal'Evedove
UFSCar

Profa. Dra. Luciana de Souza Gracioso
UFSCar

Prof. Dr. Guilherme Ataide Dias
UFPB

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Paula Regina Dal'Evedove Luciana de Souza Gracioso, Guilherme Ataide Dias e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Paula Regina Dal'Evedove

AGRADECIMENTOS

Ao único que é digno de receber todo louvor e adoração, Deus todo poderoso.

A minha mãe Maria de Jesus, ao meu irmão Dérick por ter me apoiado, dado forças e por terem acreditado neste sonho e aos meus irmãos Johnson e Calebe (*in memoriam*) na qual a saudade me fez manter o foco e a fé.

Ao meu mozão Janailton, que tem se tornado um parceiro para todas as horas.

Aos meus professores do Curso de Biblioteconomia da UFMA, que com muita dedicação apresentaram a área, os caminhos decorrentes e que embasaram o meu arcabouço para pesquisa científica.

A minha igreja Assembleia de Deus área 35 Monte das Oliveiras, por todo o apoio, oração e cuidado que tiveram comigo e com a minha família, e em especial ao Vocal Berço de Luz, ao meu Pastor Charles Nascimento, Missionária Flávia, Isaque, Samuel, tia Adriana, Tio Yan, tio Welligton, Josué, Danny Lobo e a todos os membros da congregação e da área.

A minha orientadora Paula Regina Dal'Evedove pelas ponderações, orientações, oportunidades, conversas e por acreditar na minha pesquisa.

A todos os professores do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de São Carlos - (PPGCI/UFSCar) pelo apoio e participação no processo de construção e desconstrução da pesquisa científica.

A instituição UFSCar e a CAPES pelo apoio ao Programa de Pós-Graduação em Ciência da Informação. Aos professores da minha banca, Luciana Gracioso e Guilherme Dias, pela inestimável contribuição e colocação sobre a minha pesquisa desenvolvida no mestrado, muito grata.

Aos meus amigos e parceiros Dulce, Djalda, Hellen, Paulo Nascimento, Mary, as corujinhas, que desde a graduação sempre torceram e apoiaram as nossas aventuras na pesquisa científica e que mesmo distante o apoio não foi menor. Aos meus amigos Sandra, Paulo George, Miguel, Cris, Eduarda Solidade, Ana Luiza, Breno Antunes, Felipe Zonaro, Luís Rozendo, Alexandre Leão, Thays e Bruno, muito obrigado pela força e pela amizade. Aos meus amigos Sara Raquel, Ronaldo, Rayanne Diniz pela oportunidade de crescer e aprender juntos.

Aos colegas da 3° turma do mestrado pelas discussões e compartilhamento de experiências. A todos aqueles que direta ou indiretamente contribuíram para a realização desta etapa. Muito obrigado, desejo muita sabedoria a todos.

RESUMO

O *Big Data* tem causado transformações em vários setores da sociedade e exigido novas abordagens para os estudos em Ciência da Informação. Essas transformações também implicam na emergência de novas formas de comunicação científica, intensificando a demanda por dados científicos, sendo a representação colaborativa da informação uma alternativa contemporânea aos problemas oriundos do avanço das tecnologias digitais. Nesta perspectiva, questionou-se de que forma os repositórios de dados científicos atuam como sistemas informacionais colaborativos mediante a participação ativa dos usuários ou grupos de pessoas na organização e tratamento dos registros científicos provenientes de atividades de investigação. De modo a contribuir com a questão, investigou-se as práticas de organização e tratamento da informação em repositórios de dados científicos, no intuito de contribuir com subsídios para a representação colaborativa de registros científicos provenientes de atividades de investigação, mediante os seguintes objetivos específicos: a) conceituar a representação colaborativa da informação nos estudos da Ciência da Informação e suas potencialidades em tempos de *Big Data*; b) explorar as diretrizes e práticas adotadas para a organização e o tratamento de dados científicos no âmbito da Rede de Repositórios de Dados Científicos do Estado de São Paulo; e apresentar e refletir as possibilidades da representação colaborativa de dados científicos na Rede de Repositórios de Dados Científicos do Estado de São Paulo, em todas as suas dimensões e perspectivas. Para tanto, conduziu-se pesquisa exploratória e descritiva mediante coleta de dados de ordem bibliográfica e por levantamento, tendo como escopo investigativo a literatura especializada de Ciência da Informação e os repositórios digitais pertencentes à rede. Apesar de algumas indicações teóricas na literatura científica apontarem a Folksonomia como uma possível estratégia para a organização e tratamento da informação no contexto do *Big Data*, constatou-se a inexistência efetiva dessa prática nos repositórios de dados que integram a rede, mesmo o sistema dispondo de infraestrutura tecnológica possível para o *tagging.* Conclui-se que a Ciência da Informação precisa investigar e incorporar, em suas práticas profissionais, parâmetros viáveis para a organização e recuperação de recursos informacionais implicados pelo *Big Data*.

Palavras-chave: Representação colaborativa da informação. Folksonomia. *Big Data*. Dados científicos. Repositório de dados. Rede de Repositórios de Dados Científicos do Estado de São Paulo.

ABSTRACT

Big Data has caused transformations in various sectors of society and has demanded new approaches for studies in Information Science. These transformations also imply the emergence of new forms of scientific communication, intensifying the demand for scientific data, with the collaborative representation of information being a contemporary alternative to the problems arising from the advancement of digital technologies. In this perspective, it was questioned how the scientific data repositories act as collaborative information systems through the active participation of users or groups of people in the organization and treatment of scientific records from research activities. In order to contribute to the issue, we investigated the practices of organization and treatment of information in scientific data repositories, in order to contribute with subsidies for the collaborative representation of scientific records from research activities, through the following specific objectives: a) conceptualize the collaborative representation of information in the studies of Information Science and its potential in times of Big Data; b) explore the guidelines and practices adopted for the organization and treatment of scientific data within the scope of the Network of Scientific Data Repositories of the State of São Paulo; and to present and reflect the possibilities of collaborative representation of scientific data in the Network of Scientific Data Repositories of the State of São Paulo, in all its dimensions and perspectives. To this end, exploratory and descriptive research was conducted through the collection of data of a bibliographic order and by survey, with the investigative scope of the specialized literature on Information Science and the digital repositories belonging to the network. Although some theoretical indications in the scientific literature point to Folksonomy as a possible strategy for the organization and treatment of information in the context of Big Data, it was found that there is no such practice in the data repositories that integrate the network, even though the system has possible technological infrastructure for tagging. It is concluded that Information Science needs to investigate and incorporate, in its professional practices, viable parameters for the organization and recovery of informational resources implied by Big Data.

Keywords: Collaborative representation of information. Folksonomy. Big data. Scientific data. Data repository. Network of Scientific Data Repositories of the state of São Paulo

RESUMEN

Big Data ha causado transformaciones en varios sectores de la sociedad y ha exigido nuevos enfoques para estudios en Ciencias de la Información. Estas transformaciones también implican la aparición de nuevas formas de comunicación científica, intensificando la demanda de datos científicos, y la representación colaborativa de la información es una alternativa contemporánea a los problemas derivados del avance de las tecnologías digitales. En esta perspectiva, se cuestiona cómo los repositorios de datos científicos actúan como sistemas de información colaborativos a través de la participación activa de usuarios o grupos de personas en la organización y el tratamiento de los registros científicos de las actividades de investigación. Para contribuir al tema, investigamos las prácticas de organización y tratamiento de la información en repositorios de datos científicos, con el fin de contribuir con subsidios para la representación colaborativa de registros científicos de actividades de investigación, a través de los siguientes objetivos específicos: a) conceptualizar la representación colaborativa de información en los estudios de Ciencias de la Información y su potencial en tiempos de Big Data; b) explorar las pautas y prácticas adoptadas para la organización y el tratamiento de datos científicos dentro del alcance de la Red de Repositorios de Datos Científicos del Estado de São Paulo; y presentar y reflejar las posibilidades de representación colaborativa de datos científicos en la Red de Repositorios de Datos Científicos del Estado de São Paulo, en todas sus dimensiones y perspectivas. Con este fin, se realizó una investigación exploratoria y descriptiva a través de la recopilación de datos de un orden bibliográfico y por encuesta, con el alcance investigativo de la literatura especializada en Ciencias de la Información y los repositorios digitales pertenecientes a la red. Aunque algunas indicaciones teóricas en la literatura científica apuntan a la folksonomía como una posible estrategia para la organización y el tratamiento de la información en el contexto de Big Data, se encontró que no existe tal práctica en los repositorios de datos que integran la red, a pesar de que el sistema tiene posible infraestructura tecnológica para etiquetado. Se concluye que la Ciencia de la Información necesita investigar e incorporar, en sus prácticas profesionales, parámetros viables para la organización y recuperación de los recursos informativos implicados por Big Data.

Palabras Clave: Representación colaborativa de la información. Folksonomía. Big Data Datos cientificos. Repositorio de datos. Red de Repositorios de Datos Científicos del Estado de São Paulo.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AGUIA - Agência USP de Gestão da Informação Acadêmica

AWI - Alfred Wegener Institute for Polar and Marine Research

BRAPCI - Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação

BV-CDI - Biblioteca Virtual do Centro de Documentação e Informação

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CGB - Coordenadoria Geral de Bibliotecas

CG-DIC - Comitê de Governança de Dados, Informação e Conhecimento

CKAN - Comprehensive Knowledge Archive Network

CL-DIC - Comitê Local de Gestão de Dados, Informação e Conhecimento

CNPTIA/EMBRAPA - Embrapa Informática Agropecuária

CRBU - Coordenadoria da Rede de Bibliotecas da Unifesp

CSV - Comma-Separated-Values

DCBD - Descoberta do Conhecimento em Bases de Dados

DCI - Departamento de Ciência da Informação

DMP - Data Management Platform

DOI - Digital Object Identifier

E-Dados - Escritório de Dados Estratégicos Institucionais da UNIFESP

FAIR - Findable, Accessible, Interoperable e Reusable

FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo

GDP - Gestão de Dados de Pesquisa

IBGE - Instituto Brasileiro de Geografia e Estatística

IFES - Instituições Federais de Ensino Superior

ITA - Instituto Tecnológico de Aeronáutica

JSON – JavaScript Object Notation

KDD - Knowledge Discovery in Databases

KOS - Knowledge Organization System

LC - Library of Congress

MARUM - Center for Marine Environmental Sciences da University of Bremen

NSB - National Science Board

NSF - National Science Foundation

NTI - Núcleo de Tecnologia de Informação

OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting

OCDE - Organisation for Economic Co-operation and Development

OCLC - Online Computer Library Center

OECD - Organisation for Economic Co-operation and Development

PANGAEA - Data Publisher for Earth & Environmental Science

PD&I - Pesquisa, Desenvolvimento e Inovação

PDDB - Base de Imagens de Sintomas de Doenças de Plantas

ProPq - Pró-Reitoria de Pesquisa

PRP - Pró-Reitoria de Pesquisa

PRPG - Pró-Reitoria de Pós-Graduação

RE3DATA - Registry of Research Data Repositories

REDU - Repositório de Dados de pesquisa da Unicamp

RI - Repositório Institucional

SIBi - Sistema Integrado de Bibliotecas

SIN - Secretaria de Inovação e Negócios

SIn - Secretaria de Informática

SIRE - Secretaria de Inteligência e Relações Estratégicas

SisBi - Sistema de Bibliotecas

SPD - Secretaria de Pesquisa e Desenvolvimento

STI - Superintendência de Tecnologia da Informação

TICs - Tecnologias de Informação e Comunicação

UFABC - Universidade Federal do ABC

UFSCar - Universidade Federal de São Carlos

UNESP - Universidade Estadual Paulista

UNICAMP - Universidade Estadual de Campinas

UNIFESP - Universidade Federal de São Paulo

URL - Uniform Resource Locator

USP - Universidade de São Paulo

W3C - World Wide Web Consortium

XML – Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização da pesquisa

O presente século é marcado tecnologicamente por homens e máquinas produzindo e consumindo uma grande quantidade de dados, informações e documentos em ambiente digital, com formatos e conteúdos variados. O surgimento da Internet e o avanço das Tecnologias de Informação e Comunicação (TICs) propiciam, dentre uma gama de possibilidades, o acesso cada vez mais facilitado e instantâneo a esses recursos informacionais disponíveis no ciberespaço e, por consequência, transformações substanciais na relação homem-informação.

Apesar de a humanidade produzir e disseminar dados e informações muito antes das tecnologias atuais, é a partir da Segunda Guerra Mundial (1936-1945) que se efetivam as primeiras preocupações com o armazenamento, a quantidade, o processamento e a exploração destes recursos estratégicos em diferentes áreas do conhecimento. Na Ciência da Informação[1], essas implicações tecnológicas impõem novos desafios como o desenvolvimento de metodologias tecnológicas e gerenciais que orientem as etapas de geração, desenvolvimento de coleção, armazenamento, análise e interpretação dos dados numa grande diversidade de contextos disciplinares (SAYÃO; SALES, 2019).

A criação de recursos informacionais digitais acontece de maneira vertiginosa desde o final do século XX, a partir da expansão tecnológica e a aceleração dos processos de transformação digital. Como resultado desse cenário nunca antes vivenciado na história da humanidade, tem-se uma massa informacional maior do que se pode medir e organizar individualmente – o fenômeno *Big Data*, instaurado na *Web* de dados. Em linhas gerais, o termo *Big Data* surge no contexto da pós-modernidade "[...] para descrever o crescimento, a disponibilidade e o uso exponencial de informações digitais estruturadas e não estruturadas", não devendo ser considerado uma nova tecnologia ou um novo conceito (BARRETO, 2014, p. 5).

[1] Nhacuonggue e Ferneda (2015, p. 8) esclarecem que "a Ciência da Informação nasce na necessidade de desenvolver estratégias para resolver problemas causados por ciências e tecnologias do clássico, impondo mudanças significativas no conhecimento, como visão técnico-sistema para visão usuário/humano". Com efeito, a Ciência da Informação torna-se um campo emergente que se consagra com a pós-modernidade, na segunda metade do século XX, principalmente com o movimento acelerado das TICs.

O *Big Data* tem causado transformações em vários âmbitos e exigido novas abordagens para os estudos em Ciência da Informação e que implicam no futuro da área (BARACHO et.al,2014; RIBEIRO, 2014). Essas transformações também implicam na emergência de novas formas de comunicação científica, intensificando a demanda por dados científicos. Este cenário é fruto da forte tendência computacional vivenciada nas últimas décadas nos processos de construção do conhecimento científico, com a possibilidade de realização de sofisticadas simulações de fenômenos complexos. Por decorrência, nos últimos anos surge um novo paradigma[2] na ciência – exploração/processamento de grandes volumes de dados (o chamado *Big Data*) e uso de métodos computacionais de alto desempenho, movimento conhecido como eScience ou "Ciência impulsionada pelos dados" (CORDEIRO et al.; 2013, BRAGHETTO; GOLDMAN; KON, 2013).

Sayão e Sales (2013, p. 2) apresentam a tecnologia como sendo "um elemento onipresente" no desenvolvimento da pesquisa científica ao passo que aumenta a capacidade dos instrumentos científicos, viabiliza a reconstrução de realidades por meio de simulação, além de favorecer as práticas de colaboração e compartilhamento de dados e informações. Com efeito, pesquisadores, instituições acadêmicas e agências de fomento à pesquisa passam a considerar "[...] que os dados, se devidamente tratados, preservados e gerenciados, podem constituir uma fonte inestimável de recursos informacionais" (SAYÃO; SALES, 2013, p. 3).

Os dados científicos[3], também denominados de dados de pesquisa ou dados de investigação abertos, recebem atenção especial nos últimos tempos por atuarem como recursos informacionais estratégicos para o avanço da pesquisa científica. Dados científicos "muito rapidamente deixam de ser meros subprodutos das atividades de pesquisa e se tornam um foco de grande interesse para todo o mundo científico", especialmente para as universidades e agências de fomento à pesquisa, as quais passam a atuar como protagonistas na produção e na disseminação do

---

[2] Segundo Jim Gray (HEY et al., 2009), a ciência nasceu há milhares de anos de forma *empírica*, descrevendo fenomenos naturais. Ao longo dos séculos passou a incorporar o componente *teórico*, seguido nas últimas décadas por uma forte tendência *computacional* que resultou na exploração de grandes quantidades de dados, tendência mundial conhecida por eScience.

[3] Nesta pesquisa adotamos o termo "dados científicos" por ser a terminologia empregada pela FAPESP para designar a Rede.

conhecimento, além de grandes produtoras e consumidoras desses recursos informacionais (SAYÃO; SALES, 2016, p. 91).

A fim de pesquisadores e instituições terem acesso aberto aos dados associados aos resultados de pesquisa e usarem as informações da ciência com maior eficiência, plataformas especializadas estão em funcionamento e/ou desenvolvimento em todo o mundo. Em geral, os repositórios de dados científicos, ambiente digital que permite reunir, organizar e disseminar os dados primários de uma pesquisa, estão sincronizados com o movimento da Ciência Aberta, atuando na gestão e na definição de políticas que estabeleçam as melhores práticas na curadoria digital de dados científicos, a fim de contribuir com uma pesquisa científica cada vez mais colaborativa, aberta e transparente, com vistas à aceleração de novas descobertas científicas.

De acordo com Walport e Brest (2011), os repositórios de dados científicos configuram-se como uma estratégia eficiente para a organização, preservação e compartilhamento dos dados científicos. No entanto, destacam que estes sistemas informacionais devem ser bem estabelecidos e disporem de ferramentas que sejam capazes de descrever e divulgar os dados científicos de modo a promover seu amplo acesso e reutilização.

1.2 Problema de pesquisa

Humanidades e tecnologias digitais favorecem o encaminhamento de reflexões conceituais e práticas inovadoras no campo de conhecimento da Ciência da Informação na busca por produtos e serviços informacionais que satisfaçam e atendam às necessidades atuais de sujeitos e instituições informacionais. Ribeiro (2014) e Santana Júnior et al (2014) defendem que o papel da Ciência da Informação no *Big Data* é prover meios para o acesso às informações relevantes, para construir teorias e desenvolver estudos relacionados para além das tecnologias digitais. Para os referidos autores, as questões a serem discutidas não se concentram mais nos aspectos tecnológicos, mas sim na organização e na representação da informação na era do *Big Data*.

O *Big Data* instaura a necessidade de um novo comportamento para lidar com recursos informacionais digitais que se apresentam de forma não linear, não hierárquica e sem fronteiras, em que os processos tradicionais para o tratamento da informação se tornam insuficientes, causando desafios e novas perspectivas de atuação, discussão e interlocução na Ciência da Informação (BARACHO et al., 2014; BARRETO, 2014; RIBEIRO, 2014; SANTANA JUNIOR et al., 2014).

O *Big Data* é um desafio para a Ciência da Informação no papel de conjugar os dados, os metadados e as informações harmoniosamente, útil e de fácil recuperação (SOUZA; ALMEIDA; BARACHO, 2013; DIAS; VIEIRA, 2013). Apesar de configurar-se como o resultado da explosão informacional vivido atualmente em todas as áreas do conhecimento, o *Big Data* também é a solução para os desafios que se colocam ao tratamento da informação no ambiente digital (PIMENTA, 2013). Isto porque a base do ecossistema de *Big Data* está em princípios e técnicas da Ciência da Informação que envolvem as atividades de coleta, representação, armazenamento e disseminação seletiva de dados e informações (BARBOSA; KOBASHI, 2017; VICTORINO et al., 2017).

Brooks (2001) menciona que o modelo clássico de representação da informação pode não ser mais aplicável ou econômico no ambiente digital, visto que o *Big Data* é um grande fenômeno heterogêneo, descentralizado e com alta taxa de crescimento. Como alternativa à questão, tem-se cada vez mais a necessidade dos usuários personalizarem seus metadados, catalogando seus dados em ambiente digital seguro, confiável e íntegro de forma colaborativa (BROOKS, 2001). Sobre esse ponto, Souza, Almeida e Baracho (2013) evidenciam o impacto do *Big Data* em técnicas tradicionais de análise de assunto, análise documentária ou análise temática, ao passo que a nova era informacional rompe com a possibilidade do tratamento individual e intelectual dos registros.

Neste sentido, dados e informações dispostos na *Web* podem ter uma vida útil mais longa se comparado ao material físico, e que combinados com o potencial para atingir os usuários dispostos na rede, sem custos incrementais significativos, podem mudar a forma de produção, organização, acesso e disseminação dos mais variados recursos informacionais. Exemplo disso é a capacidade revolucionária de comunicação, colaboração e compartilhamento de informações digitais entre milhares

de usuários, sendo ao mesmo tempo autores, editores, disseminadores e indexadores das informações (QUINTARELLI, 2005).

Hjorland (2012) e Ibekwe-Sanjuan e Bowker (2017) esclarecem que o *Big Data* e o Google não removeram a necessidade de classificação, pelo contrário, com o alto volume e tipos de dados há uma necessidade de classificação maior do que em outros tempos, principalmente a classificação feita por seres humanos. Uma vez que não, necessariamente, é um algoritmo que determina os critérios de classificação, esta pode ser feita de maneira mais adequada por pessoas.

Além disso, na classificação automática pode ser difícil obter um nível de qualidade suficiente. De modo complementar, os referidos autores sinalizam que as dinâmicas associadas ao *Big Data* são mais otimizadas com bons metadados, que o mesmo aumentou as tarefas em muitos campos e que se baseia na coexistência de duas modalidades aparentemente opostas – trabalho humano e automação (HJORLAND, 2012; IBEKWE-SANJUAN, BOWKER, 2017).

Ao considerarmos que o *Big Data* só funciona com bons metadados, pois a incorreta criação, manipulação e uso podem gerar dados discriminatórios e encaminhar a pesquisa há resultados não confiáveis, é imprescindível que o campo da informação se debruce cada vez mais para as classificações facetadas de abordagens *top down* ao invés da aristotélica. Com efeito, é preciso atingir o equilíbrio certo entre duas abordagens opostas – incorporar sistemas de classificação construídos por especialistas e solicitar contribuições de amadores via plataforma Web 2.0 para a etiquetagem colaborativa dos recursos informacionais presentes em ambientes digitais (HJORLAND, 2012; IBEKWE-SANJUAN, BOWKER, 2017).

Moura (2009, p. 61) evidencia que “a Web 2.0 privilegiou a efetiva colaboração do usuário no processo de organização e personalização da informação [...]”, assim como rompeu com a lógica de organização da informação, antes centralizada no papel do mediador/gestor da informação. Por decorrência, o usuário passa a exercer um papel cada vez mais relevante na criação, tratamento, recuperação, disseminação, acesso e manutenção de recursos informacionais, nos mais variados ambientes digitais e para os mais diferentes propósitos.

A partir dessas premissas, questiona-se: de que forma os repositórios de dados científicos atuam como sistemas informacionais colaborativos mediante a participação

ativa dos usuários ou grupos de pessoas na organização e tratamento dos registros científicos provenientes de atividades de investigação?

1.3 Objetivos

O objetivo geral da pesquisa é investigar as práticas de organização e tratamento da informação em repositórios de dados científicos, no intuito de contribuir com subsídios para a representação colaborativa de registros científicos provenientes de atividades de investigação.

Para tanto, são delineados os seguintes objetivos específicos:

- Conceituar a representação colaborativa da informação nos estudos da Ciência da Informação e suas potencialidades em tempos de *Big Data*;
- Explorar as diretrizes e práticas adotadas para a organização e o tratamento de dados científicos no âmbito da Rede de Repositórios de Dados Científicos do Estado de São Paulo;
- Apresentar e refletir as possibilidades da representação colaborativa de dados científicos na Rede de Repositórios de Dados Científicos do Estado de São Paulo, em todas as suas dimensões e perspectivas.

1.4 Justificativa

A presente pesquisa surge de um contexto social, tecnológico e científico amplo que a justifica, e se materializa em apontamentos e discussões apresentadas sobre as práticas de representação colaborativa da informação em tempos de *Big Data*, tendo os dados científicos como escopo investigativo.

O volume, a velocidade e a variedade de formatos e conteúdos de dados e informações no ambiente digital crescem de forma vertiginosa. Isto implica no desenvolvimento de práticas e soluções inovadoras que garantem o acesso e a recuperação de recursos informacionais digitais confiáveis e representativos, em atenção à demanda dos usuários de sistemas de recuperação de informações. Para

além dos seus benefícios, o avanço das tecnologias digitais e o uso cada vez mais intenso do ciberespaço culminou no excesso de informações fragmentadas e, por decorrência, em problemas e grandes transtornos de acesso às fontes e conteúdos informacionais precisos. Esta realidade também é observada quando consideramos os dados científicos enquanto recursos informacionais disponíveis em plataformas de acesso aberto e enquanto objetos de pesquisa recentes na Ciência da Informação.

Ao passo que o *Big Data* causa implicações de acesso e recuperação dos recursos informacionais digitais, atribuir metadados precisos, confiáveis e representativos aos dados científicos requer o desenvolvimento de sistemas informacionais colaborativos que viabilizem a integração dos usuários na representação desses recursos com fins de compartilhamento e recuperação. Além disso, a representação colaborativa de dados científicos potencializa a interoperabilidade, assim como tem feito a *Library of Congress* (LC) na condição de participante da *Online Computer Library Center* (OCLC) que disponibiliza seu catálogo de imagens no WorldCat para o tagueamento realizado por usuários, que vincula esses metadados ao seus vocabulários controlados.

Em estudo conduzido por Mathes (2004), identificou-se na rede social Flickr que os usuários, ao atribuírem *tags* a uma imagem, registravam as especificações da câmera, como tipo, marca, ano, além de termos que representava o conteúdo da imagem. Na Folksonomia, os usuários representam a informação e/ou os metadados de forma a indicar o conteúdo, assim como podem atribuir termos que descrevem a estrutura digital e as caraterísticas específicas do recurso informacional em análise. Sendo assim, considera-se estar relacionado e imbricado na Folksonomia os dois tipos de representação da informação, a temática e a descritiva no contexto da *Web* social para garantir a organização e a recuperação dos dados científicos.

De modo complementar, o acesso aos dados científicos de forma precisa, útil, rápida e relevante torna-se crucial na conjectura atual de *Big Data*, sendo confluência destes dois tipos de representação fundamental. Na visão de Trant (2008), a marcação ou as *tags* surgem como uma possível solução para o processo de busca de recursos informacionais em rede, bem como meios para apoiar seu uso personalizado.

Nos ambientes folksonômicos predomina-se a ação do usuário de atribuir termos representativos aos recursos informacionais, com autonomia de escolha livre

do vocabulário utilizado para etiquetar. Outro fator importante a ser incluído é o número grande e variado de usuários etiquetando os mesmos recursos informacionais, cujos tagueamentos revelam palavras-chave do descritor de assunto em vários níveis de especificidade, abordando assuntos técnicos, questões de gênero, temas gerais, nome de lugares, anos, neologismos, responsáveis pelos recursos, dentre outros (MATHES, 2004).

Assim como em sistemas de informação, nestas redes sociais é possível identificar dois tipos de registros feitos pelo usuário da rede: o registro bibliográfico, com a representação temática e descritiva do documento, e o registro de identidade que se refere aos responsáveis. Depreende-se que é possível potencializar a organização e a recuperação dos dados científicos com a diminuição do controle dos termos de acesso, pois nestas redes sociais os termos são indexados por várias pessoas, cujo resultado do processo se torna mais exaustivo, tem maior número e variedade de termos, maior consistência e com a linguagem natural dos usuários.

A Folksonomia se apresenta como um modelo diferenciado de representação de assunto, em que os próprios usuários analisam o conteúdo de um determinado recurso informacional e assimilam termos descritivos (*tags*), gerando um índice do qual pode-se recuperar vários outros. Neste sentido, importa considerar que a cada dia a coleção de recursos informacionais dispostos no ambiente digital cresce exponencialmente, situação que reforça a necessidade dos metadados para realizar uma descrição estruturada dos atributos essenciais dos dados científicos de maneira rápida e eficiente (GILL, 2008).

Apesar de a Folksonomia não ser uma solução universal para o problema da organização e recuperação da informação precisa e satisfatória no ambiente digital, a seleção do conjunto mais apropriado de ferramentas e padrões de metadados permitirá descrições mais completas e contextualizadas dos recursos informacionais, bem como o mapeamento de metadados criados de acordo com diferentes comunidades, disciplinas, padrões específicos, aumentando a interoperabilidade e garantindo a abrangência dos termos descritivos (CATARINO e BAPTISTA, 2007; GILLILAND, 2016).

Dessa maneira, é central para o esforço científico adaptar-se às novas temporalidades do desenvolvimento teórico e técnico ocasionado pelo *Big Data* na Ciência da Informação (HJORLAND, 2012, 2013; IBEKWE-SANJUAN, BOWKER,

2017). Nessa perspectiva, segundo Oliveira et al. (2017) a Folksonomia ganha cada vez mais espaço por ser uma forma de classificação que não utiliza taxonomias ou vocabulários pré-estabelecidos, não há restrição hierárquica pré-definida ou restrição para definir as possíveis *tags*. Diferentemente, as taxonomias formais e ontologias são baseadas em estrutura hierárquica, sua manutenção é dificultosa, seus metadados são gerados por *experts*, seu vocabulário é controlado, requer acordo consensual sobre os seus conteúdos e a sua representação exige mais recursos para a criação e manutenção dos sistemas. Por outro lado, segundo Oliveira et al. (2017, sem paginação) a Folksonomia cria *namespaces* planos e abertos aos usuários para editarem as *tags*, é de fácil manutenção, não requer acordo consensual, quanto aos seus metadados podem ser gerados por *experts* e pelos usuários da informação, as palavras-chaves são escolhidas de forma livre, sem uso de vocabulários controlados.

Deve-se destacar ainda que as práticas de representação colaborativa da informação em repositórios de dados científicos é um tema escasso na literatura especializada, sendo oportuno a condução de discussões teóricas e aplicadas que tragam subsídios ao campo de conhecimento da Ciência da Informação e contribua com novas frentes de investigação. No Grupo de Pesquisa Representação e Humanidades Digitais da Universidade Federal de São Carlos, o tema é investigado no âmbito do Projeto de Pesquisa “Representação e recuperação por assuntos em repositórios institucionais”, sob coordenação da Profa. Da. Paula Regina Dal’Evedove – Professora Permanente do Programa de Pós-Graduação em Ciência da Informação desta mesma instituição. Nesta perspectiva, a pesquisa corrobora com o referido projeto de pesquisa, tendo como escopo investigativo a representação colaborativa de dados científicos em tempos de *Big Data*.

A partir do exposto, considera-se pertinente a condução de estudos que investiguem as iniciativas de representação colaborativa aplicada aos dados científicos no âmbito dos repositórios de dados científicos brasileiros, bem como contribuir para o avanço do tema na literatura nacional e internacional de Ciência da Informação. Desta forma os dados científicos se contextualizam como um produto do *Big Data* e os repositórios como uma infraestrutura eficiente e eficaz que consegue abarcar esses dados no contexto atual.

1.5 Percurso metodológico

Em virtude da complexidade do Universo, cabe ao pesquisador encontrar formas para organizar o conhecimento para descobrir os fenômenos (CINTRA, 1982). Para a realização desta pesquisa um rol de procedimentos foi adotado a fim de descortinar as práticas de organização e tratamento de dados científicos na perspectiva do *Big Data* e da literatura científica nacional e internacional de Ciência da Informação. De igual forma, a pesquisa se caracteriza como um estudo de caso na Rede de Repositórios de Dados Científicos do Estado de São Paulo com o intuito de identificar como os dados científicos estão sendo representados nesta plataforma de acesso aberto.

Trata-se também de uma pesquisa exploratória e descritiva, com procedimentos de análise de conteúdo de um *corpus* específico. Este método permite o estudo do tema sob diversos ângulos e aspectos, considerando que tem como finalidade proporcionar mais informação sobre o assunto que se propõem a investigar, possibilitando sua definição e delineamento (SILVA, 2004).

De natureza qualitativa procura aprofundar-se na compreensão dos fenômenos propostos, sendo considerada bibliográfica por recorrer ao levantamento do referencial teórico, utilizando fundamentalmente das contribuições de vários autores sobre os temas supracitados. Segundo Gil (2008), a pesquisa bibliográfica permite a cobertura de uma gama de fenômenos mais amplos do que aquela que se poderia pesquisar diretamente, sendo consideradas publicações científicas nos idiomas português, espanhol e inglês. Dentre as vantagens do uso de diversas fontes documentais, tem-se o conhecimento do passado, a investigação dos processos de mudança social e cultural, a obtenção de dados com menor custo, além de favorecem a obtenção de dados sem o constrangimento dos sujeitos e permitir resultados mais acurados (GIL, 2008).

As fontes de informação utilizadas para o embasamento da pesquisa foram oriundas de um extenso levantamento bibliográfico realizado na Ciência da Informação que versam sobre *Big Data*, dados científicos, folksonomia e repositórios de dados. Na investigação optou-se por fazer o levantamento das pesquisas científicas na Base de Dados Referencial de Artigos de Periódicos em Ciência da

Informação (BRAPCI) e na base de dados PERI- ECI, as quais indexam artigos de periódicos e trabalhos publicados em anais de eventos técnico-científicos da área de Ciência da Informação, assim como no Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

No sentido de apresentar um panorama de discussão teórica sobre a representação colaborativa da informação em tempos de *Big Data*, foram empregados nos campos assunto, título, resumo e palavras-chave nas referidas bases de dados os seguintes termos de busca: “representação colaborativa”, “folksonomia ”, “indexação social”, “classificação social”, “classificação popular”, “etiquetagem colaborativa”, “indexação colaborativa”, “Web colaborativa”, “tagueamento social”, “representação colaborativa da informação” e “tagging”. O limite temporal das buscas foi os textos científicos publicados até o ano de 2019, embora não tenha delimitado o tempo de início os textos recuperados são datados a partir de 2008. Deu-se preferência pelos materiais publicados em língua portuguesa do Brasil, para identificar esta discussão no contexto brasileiro, e algumas publicações em língua Inglesa.

Após a recuperação do material nas bases de dados, procedeu-se com a leitura dos resumos, introdução, métodos e considerações finais de modo a identificar e selecionar os materiais em que a folksonomia tenha sido objeto de estudo principal. Após esta seleção dos materiais retornados na busca, visando à análise qualitativa dos dados coletados, realizou-se a leitura integral do material e a elaboração de síntese teórica em torno dos principais aspectos abordados.

Godoy (1995, p. 22) esclarece que para entender a dinâmica do fenômeno em estudo é necessário a captação do mesmo a partir da perspectiva das pessoas envolvidas. No contexto da presente pesquisa, essa captação do fenômeno ocorreu por meio de documentos – tipificados na ordem de artigos científicos, teses, dissertações, planos de gerenciamento de dados, políticas de submissão, políticas de autoarquivamento, resoluções –, vislumbrando-se nos mesmos a qualidade de fontes potencialmente ricas em dados e que “podem ser considerados uma fonte natural de informação à medida que, por terem origem num determinado contexto histórico, econômico e social, retratam e fornecem dados sobre esse mesmo contexto.”

Para responder ao problema de pesquisa “de que forma os repositórios de dados científicos atuam como sistemas informacionais colaborativos mediante a participação ativa dos usuários ou grupos de pessoas na organização e tratamento

dos registros científicos provenientes de atividades de investigação?", procedeu-se a coleta de dados sobre a Rede de Repositório de Dados Científicos do Estado de São Paulo enquanto ferramenta desenvolvida pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), com apoio de universidades paulistas, utilizando como fonte de informações os sítios eletrônicos da própria instituição, o portal de acesso ao repositório, os repositórios de dados associados, os *softwares*, os planos de gerenciamento de dados e outros documentos que discorriam sobre a plataforma publicados ou não pela instituição. Na situação de indisponibilidade de informações nos sítios eletrônicos, encaminhou-se e-mail para os gestores dos Sistemas de Bibliotecas, que são os responsáveis pelos repositórios de dados científicos, solicitando-as, porém foram poucas as respostas retornadas. Os dados foram coletados no período de dezembro de 2019 a março de 2020.

Como instrumento auxiliar para a análise dos dados, utilizamos o conjunto de sete categorias básicas propostas por Tomaél e Silva (2007) para análise de repositórios institucionais, as quais ajudam a ter um ponto de partida estruturado para analisar e organizar as informações, quais sejam*:* responsabilidade, conteúdo, aspectos legais, padrões, preservação digital, política de acesso e uso, sustentabilidade e financiamento. No Quadro 1 são descritas as categorias e os respectivos aspectos abordados, conforme segue:

Quadro 1. Categorias básicas para análise dos repositórios de dados

CATEGORIAS	ASPECTOS ABORDADOS
Responsabilidade	Profissionais e instituição responsável pela gestão do repositório e como os mesmos organizam e fazem a manutenção dos dados.
Conteúdo	Conteúdo permitido no repositório, responsabilidade de fazer o depósito, formatos de arquivo aceitos, estratégias de captação de dados.
Aspectos legais	Licenças utilizadas para a disponibilização dos dados de pesquisa.
Padrões	Padrões de interoperabilidade, de metadados, de tecnologias, de fluxos de trabalhos, de interface e usabilidade.
Preservação digital	Existência de políticas de preservação de preservação digital e aspectos contemplados, e identificadores digitais.
Política de acesso e uso	Permissões de acesso e uso dos dados de pesquisa. Suporte ao usuário que submete e ao que usa os dados.

Sustentabilidade e Financiamento	Estratégias de financiamento e sustentabilidade do repositório de dados.

Fonte: adaptado de Tomaél e Silva (2007)

Nesta pesquisa consideramos estas categorias como insumo básico acrescidas de outras informações, levando em consideração os repositórios de dados científicos como objetos de estudo. Além deste instrumento, foram elaborados roteiros como instrumentos auxiliares na identificação e análise das políticas e do ambiente digital da Rede de Repositório de Dados Científicos da FAPESP, de modo a identificar como os dados científicos estão sendo representados e identificar ações de folksonomia conduzidas no contexto da plataforma para o tratamento dos registros científicos provenientes de atividades de investigação, descritos nos Quadros 2 e 3.

Considerando o levantamento de informações necessárias para a identificação de políticas da rede, conhecimento dos seus processos de representação de assunto, identificação de adoção de formas colaborativas, tanto nas políticas, quanto nos sites dos repositórios, elaboramos perguntas que auxiliassem na busca destas informações. Organizamos estas perguntas nos Quadros 2 e 3, descritos abaixo.

Quadro 2. Roteiro para análise das políticas da Rede FAPESP

ROTEIRO PARA ANÁLISE DAS POLÍTICAS DA REDE
De que forma as políticas abordam a questão dos dados científicos?
A análise e representação de assunto é contemplada nas políticas?
As políticas contemplam a representação colaborativa de dados científicos?
Quais as opções de folksonomia contempladas nas políticas?

Fonte: elaborado pela autora

Quadro 3. Roteiro para análise do ambiente digital do repositório

ROTEIRO PARA ANÁLISE DO AMBIENTE DIGITAL DO REPOSITÓRIO	
Caso haja a possibilidade de adicionar tags	**Caso NÃO haja a possibilidade de adicionar tags**
Há um campo visível para o usuário adicionar termos/*tags*?	Como é realizada a representação dos dados científicos?
Como é realizada a adição de *tags*?	Quem está autorizado a descrever os conteúdos informacionais dos dados científicos?
Quem é autorizado a adicionar *tags*?	

Fonte: elaborado pela autora

Após o levantamento e leitura da bibliografia sobre os temas abordados nesta pesquisa, deu-se a escrita dos capítulos desta dissertação. A seguir identificamos a Rede de Repositório de Dados Científicos da FAPESP e as instituições participantes, conhecendo seus sites, repositórios e políticas.

Quanto a coleta de informações e análise da Rede, foi realizado um levantamento nos sites da FAPESP e das 8 instituições participantes, uma a uma, de políticas, normas, resoluções, manuais, planos de gerenciamento de dados, planos estratégicos, e outros documentos relacionados, publicados ou não pela instituição pesquisada.

Quanto a estes documentos, foi realizada a análise de conteúdo buscando identificar as categorias básicas proposta por Tomaél e Silva (2007) para os repositórios e as diretrizes que respaldam a inserção e uso da representação colaborativa no repositório, de acordo com um roteiro previamente elaborado (Quadro 2).

Em seguida, analisamos o ambiente digital dos repositórios de cada instituição e do portal da FAPESP que conjuga as informações de todos, de acordo com um roteiro elaborado para a pesquisa (Quadro 3).

Ao término deste processo, realizamos discussões, considerações e ponderações sobre algumas possibilidades de sistematizar a representação colaborativa de dados científicos na Rede de Repositórios de Dados Científicos do Estado de São Paulo.

1.6 Estrutura da pesquisa

Para a apresentação, desenvolvimento e alcance dos objetivos propostos para esta pesquisa, este trabalho está subdividido em cinco capítulos, a saber:

Capítulo 1: Introdução – descrevem-se os temas gerais e os fundamentos que foram abordados no desenvolvimento desta pesquisa, o problema, as justificativas de investigação, os objetivos, o percurso metodológico e a estrutura formal do conteúdo desta dissertação.

Capítulo 2: O fenômeno *Big Data* – apresentam-se as bases teóricas e conceituais do *Big Data* e as implicações deste na Ciência da Informação, sobretudo, as implicações na Organização do Conhecimento que dialogam e enriquecem o estudo. Também abordamos as concepções teóricas e fundamentos dos Dados científicos e dos Repositórios de dados. Nesta seção apresentamos o *Big Data* como um fenômeno que perpassa a área da Ciência da Informação, os dados científicos como produto deste fenômeno e os repositórios como a infraestrutura ideal para promover a organização, gestão, indexação, recuperação e compartilhamento dos dados.

Capítulo 3: Representação colaborativa da informação em ambientes digitais – discorre-se sobre os aspectos conceituais, epistemológicos e sistêmicos da representação colaborativa, utilizando dos principais pesquisadores, nacionais e internacionais, que estudam a temática na área. Neste capítulo apresentamos a representação colaborativa como uma metodologia ou proposta viável para indexar, classificar e recuperar os dados científicos disponíveis nos repositórios de dados, no contexto do *Big Data* de forma a auxiliar no processo de organização da informação. Por conseguinte, apresentamos os principais estudos mapeados na literatura sobre a representação colaborativa da informação na perspectiva do *Big Data*.

Capítulo 4: Análise da rede de repositórios de dados científicos do estado de São Paulo - apresentação dos resultados, análises e descrição das políticas, do ambiente digital e das categorias básicas dos repositórios de dados das instituições.

Capítulo 5: Considerações finais – apresentam-se algumas reflexões sobre os objetivos iniciais propostos por esta pesquisa, sobre os resultados alcançados, bem como sugerem-se possibilidades para futuras investigações, o apontamento de melhorias e investigação sobre a temática em outros repositórios, seja nacionais ou internacionais.

2 O FENÔMENO *BIG DATA*

A sociedade do século XXI é marcada pelo acesso facilitado e rápido à informação, possibilitado pelo desenvolvimento avançado e acelerado de tecnologias que permitem o acesso e troca de dados, a exemplo de computadores, *smartphones*, *tablets*, Internet, *Smart* TV, se comparado com séculos passados. Segundo o relatório da Conferência das Nações Unidas sobre Comércio e Desenvolvimento (2017), o Brasil é o 4° maior país com maior número absoluto de usuários de Internet do mundo. O Instituto Brasileiro de Geografia e Estatística (IBGE) identificou que em 2015 o número de usuários de Internet no Brasil era de 116 milhões, sendo que do total de domicílios (69,3 milhões), em 45,3% existe microcomputador, 92,6% *smartphones* e que em 70% existe o acesso à Internet.

Com o grande número de acessos a estas tecnologias de informação e comunicação no Brasil e no mundo, influenciado por alguns fatores como diminuição de custo financeiro, aumento da capacidade de processamento e a diminuição física do *hardware*, a produção de dados, informações, conhecimentos e documentos diversos foi potencializada, elevando-se a níveis exponenciais jamais vistos em outras épocas da sociedade. Outros fatores que influenciaram nesta potencialização foram as mudanças das abordagens *top-down* para *bottom-up* em alguns ambientes, principalmente os digitais. A democratização do acesso, a produção de conteúdo e a disseminação por parte dos usuários também auxiliaram neste processo.

Este cenário de grande produção de dados e informações decorre desde o final do século XX com a rápida expansão das tecnologias de informação e comunicação e a aceleração dos processos de transformação da mesma (criação, representação, armazenamento, organização, disseminação e consumo). Apesar de a humanidade produzir e disseminar informação muito antes das tecnologias atuais surgirem, é a partir da Máquina de Turing, vale dizer, os computadores, desenvolvida pelo professor de matemática teórica e aplicada Alan Turing durante a Segunda Guerra Mundial (1936), hoje reconhecido como o pai da Ciência da Computação, que se iniciam as primeiras preocupações com os dados, seu armazenamento, sua quantidade, seu processamento e a exploração destes.

Como contextualizado anteriormente, o *Big Data* cresce exponencialmente a cada dia, pois os dados são produzidos tanto de forma intencional como não intencional. A exemplo de um dado produzido de forma intencional, quando um

usuário acessa uma rede social, publica uma foto e adiciona *hashtags*[4]. Não intencional, quando este usuário se locomove de um ponto a outro e seu *smartphone* registra automaticamente a localização e os estabelecimentos próximos, a partir desta situação esse usuário começa a receber propagandas dos estabelecimentos e sugestões de lugares em seu *smartphone,* email e redes sociais no geral.

Ao acessar redes sociais, aplicativos de geolocalização, conteúdos em vídeo, comércio eletrônico, dados públicos e outros conteúdos na Web, os serem humanos produzem pegadas digitais, dados estruturados e não estruturados em quantidade volumosa. Portanto, a essência do *Big Data* é coletar esta grande quantidade de dados produzida por nossas interações constantes com serviços e dispositivos de interação e comunicação e transformá-la em conhecimento (FOX; HENDLER, 2011). Nesse contexto, o *Big Data* se refere à gestão de complexos conjuntos de dados, com o intuito de evidenciar por meio de processos analíticos associados tendências e conexões com potencial de contribuir para a estratégia de negócios (GANDOMI; HAIDER, 2015).

Isaac Asimov (1988) em entrevista ao programa de *TV World of Ideas* predizia que para se ter sucesso era necessário aprender com a informação, e as empresas e organizações estão estrategicamente explorando o *Big Data* para alcançá-lo. Na concepção de Motta, Barbosa e Barbosa (2019, p. 86) o *Big Data* representa “[...] uma estratégia de gestão informacional que influencia o ecossistema organizacional, transforma processos gerenciais e estimula a inovação”. O *Big Data* envolve tanto a adoção de sistemas tecnológicos avançados como o desenvolvimento de habilidades intelectuais de alto nível para coletar, estocar, organizar, extrair, analisar e distribuir dados (GANDOMI; HAIDER, 2015).

O *Big Data* possui algumas acepções conhecidas como os 5Vs, descritos como Volume, Variedade, Velocidade, Veracidade e Valor (LANEY, 2012; STONEBRAKER, 2012; BARRETO, 2014). Estas características representam Volume como uma grande quantidade de dados disponíveis e produzidos paralelamente, Variedade como várias fontes e formatos em que os dados se encontram e são produzidos, Velocidade como a produção de dados constantes e ininterruptos, Veracidade como a qualidade dos

[4] Hashtags são palavras-chave ou termos associados a uma informação, tópico ou discussão que se deseja indexar de forma explícita nas redes sociais. (WIKIPÉDIA, 2020)

dados e Valor como a capacidade de produzir um conhecimento a partir da análise destes dados (LANEY, 2012; BARRETO, 2014), conforme disposto a seguir:

Quadro 4. Dimensões do Fenômeno *Big Data*

Volume	Magnitude dos dados. Grande número de dados estocados ou entrada de grande número de registros no sistema.
Velocidade	Frequência ou velocidade de geração e entrega dos dados. Capacidade de processamento em tempo real.
Variedade	Integração dos dados. Eles são gerados a partir de diversas fontes e em diferentes formatos.
Valor	Importância de se extrair benefícios institucionais e econômicos da análise dos dados.
Veracidade	Relevância da qualidade dos dados, em termos de confiabilidade das fontes e dos dados delas extraídos.

Fonte: Adaptado de Wamba et al. (2015)

Tendo em vista que o *Big Data* é derivado da ampla difusão e adoção de plataformas virtuais, mídias móveis, redes sociais e de conceitos relacionados à Internet das Coisas, Wamba et al (2015, p. 235) mencionam que “[...] nós definimos *Big Data* como uma abordagem holística para gerenciar, processar e analisar cinco Vs, de forma a criar *insights* acionáveis para a entrega sustentável de valor”. De forma complementar, o *Big Data* pode ser compreendido em uma perspectiva macro como sendo um elo que permite conectar o mundo físico e o ciberespaço. O fenômeno associa-se a uma complexa estrutura de *hardware*, *software*, objetos em rede (que são criados por dispositivos com tecnologia digital embarcada que fornecem informações sobre os usuários e contexto de uso) e habilidade intelectual (JIN et al., 2015).

O *Big Data* é composto por dados semi-estruturados, não-estruturados e dados estruturados, codificados em linguagem computacional. Gandomi e Haider (2015) explicitam que os dados não-estruturados consistem em imagens, vídeos, áudios, textos de redes sociais e de páginas da Internet, produzidos em linguagem natural. Os semi-estruturados, por sua vez, são os que adotam linguagem textual para troca de dados na Web, com *tags* definidas pelos usuários que tornam possível sua leitura por máquinas. Os dados estruturados, constituem apenas 5% de todos os dados

existentes, referem-se aos dados tabulares encontrados em planilhas ou bancos de dados relacionais.

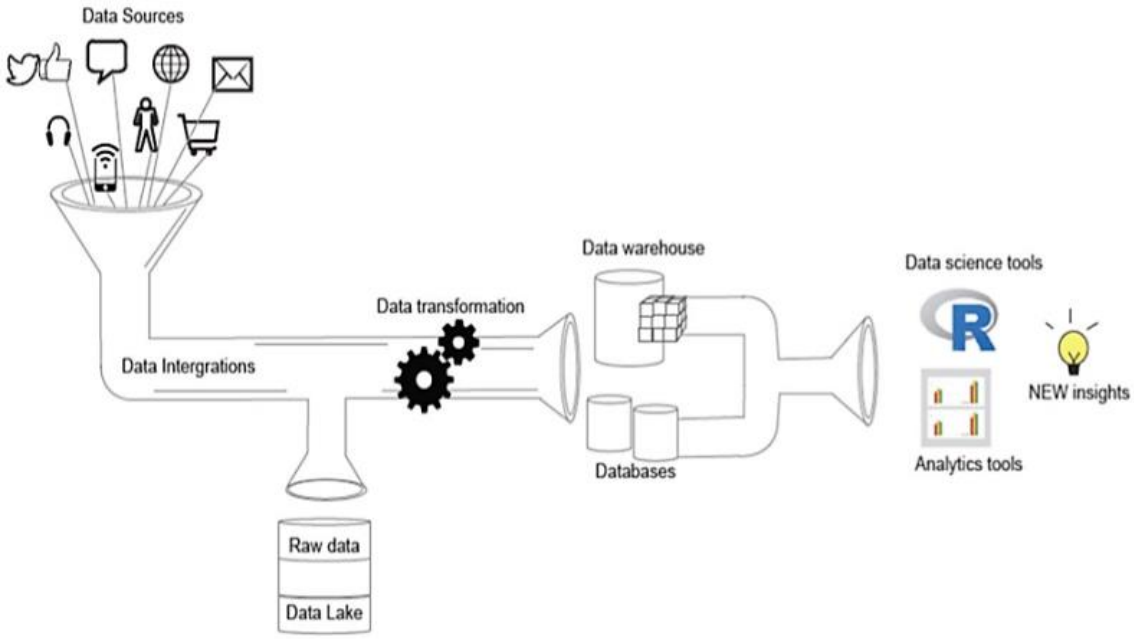
> Um alto nível de variedade, característica distintiva do *Big Data*, não é necessariamente novo. Organizações têm obtido dados não-estruturados de fontes internas (exemplo: sensores) e de fontes externas (exemplo: mídias sociais). Contudo, a emergência de novas tecnologias de gestão e análise de dados, que habilitam as organizações a aplicar dados em seus processos de negócios, é o aspecto inovador (GANDOMI; HAIDER, 2015, p. 138).

Com estas características e diferentes dados, o *Big Data* mostra a necessidade de aprender com estes para gerar informação, de encontrar informação útil entre tantas outras demasiadas. O *Big Data* possibilita a coleta e descrição dos dados, a avaliação destes e a utilização dos mesmos para prever tendências, eventos futuros e informações que agregue valor.

Uma das iniciativas pioneiras neste processo de recorrer à análise de dados para a descoberta de um novo tipo de conhecimento foi o *Knowledge Discovery* in *Databases* (KDD). O processo de Descoberta do Conhecimento em Bases de Dados (DCBD) do inglês KDD, representa uma forma de extração de novos conhecimentos contidos em grandes bancos de dados. Este processo passa por várias etapas até se chegar à possibilidade de descoberta de algo relevante que possa se tornar um conhecimento. Romão (2002) explica que esse termo teria surgido no primeiro *workshop* de KDD em 1989. A DCBD pode ser explorada por várias áreas do conhecimento, pois há dados que podem ser processados para as mais diversas finalidades. Sobre isto, Bogorny (2003, p.12) indica que “[...] as aplicações de DCBD integram teorias, métodos e algoritmos provenientes destas diferentes áreas, tendo como objetivo a extração de conhecimento a partir de grandes Bases de Dados”.

Na perspectiva de análise e gestão de *Big Data*, o processo de extração de conhecimento de um conjunto de dados perpassa por algumas etapas como seleção, pré-processamento, conversão com o formato compatível com o *software*, mineração de dados e interpretação e validação. O processo de *Big Data* tem início com a entrada de dados de diversas fontes, em seguida sistemas avançados de *hardware* e *software* são integrados, transformados e separados. Adiante, os dados são analisados por meio de ferramentas que permitem estabelecer relações, tendências e padrões de comportamento, conforme ilustrado na Figura 1.

Figura 1. Processo de gestão e análise do *Big Data*



Fonte: Daish (2017)

O KDD, assim como o *Data Mining*, *Data Warehouse*, *Data Mart*, *Business Intelligence* e o *Cloud Computing* são consideradas tecnologias bases para o *Big Data*. Mas, este fenômeno não está relacionado apenas ao aspecto tecnológico, incluem-se neste âmbito questões quanto à análise e à mitologia. O primeiro aspecto refere-se à maximização do poder de computação e precisão para coletar, analisar, vincular e comparar grande conjunto de dados. Por sua vez, o segundo aspecto é a análise de dados sendo feita por meio da execução de mineração que passa pela extração e análise de grande volume de dados em busca de padrões e comportamentos, é a essência do *Big Data*. Motta, Barbosa e Barbosa (2019, p. 88) apontam que “a análise refere-se ao uso de técnicas para extrair inteligência deles. Ela cria sentido para os dados, estabelecendo conexões e embasando soluções criativas de visualização e compartilhamento de informações”.

O aspecto da mitologia é abordado por Boyd e Crawford (2012), autores que relacionam este aspecto à crença que grande conjunto de dados vão oferecer uma forma superior de inteligência e competitividade que pode gerar insight. Este último fator é um dos motores de propulsão para o estudo e o uso de *Big Data* por grandes

corporações que têm impactado na geração de novos produtos, novas possibilidades de atuação no mundo dos negócios, na gestão e predição de cenários (LANEY, 2012).

Levando em consideração que o *Big Data* está intrinsecamente relacionado às tecnologias, as discussões nas diversas áreas do conhecimento se afloram sobre o seu conceito e suas implicações. Dumbill (2012, p. 3, tradução nossa) explica que "*Big Data* são dados que excedem a capacidade de processamento de sistemas de bancos de dados convencionais. Os dados são muito volumosos, se movem muito rápido, ou não se encaixam nas estruturas das arquiteturas de banco de dados." O *Big Data* é um corte horizontal que transpassa várias áreas do conhecimento e o seu real valor pode estar em gerar conhecimento a partir da análise do grande volume de dados caracterizados nos 5Vs. Podemos considerar, de igual forma, que o valor do *Big Data* também está na capacidade das organizações em gerenciar os dados gerados.

Boyd e Crawford (2012) e Soares (2018) apresentam a dualidade do *Big Data*, primeiro visto como ferramenta poderosa para abordar vários males da sociedade, oferecendo o potencial de novos insights sobre áreas diversas, e por outro lado visto como uma manifestação preocupante, permitindo invasões de privacidade, diminuindo liberdades dos civis e aumentando o controle por parte do Estado e empresas.

Considerando que mencionamos anteriormente que o *Big Data* é um tema de discussão em várias áreas do conhecimento, identificamos a necessidade de estudar os aspectos deste na Ciência da Informação, de identificar nos autores da área quais as perspectivas e temáticas de estudo a respeito do fenômeno na literatura científica da Ciência da Informação. Portanto, na próxima seção apresentamos e discorremos sobre as perspectivas de estudos sobre *Big Data* recuperados na literatura científica da Ciência da Informação brasileira.

2.1 Perspectivas do *Big Data* na Ciência da Informação

Em Ciência da Informação, os primeiros estudos relacionados ao *Big Data* despontam pelos anos de 2013. Estes estudos (SOUZA; ALMEIDA; BARACHO, 2013; DIAS; VIEIRA, 2013 e PIMENTA, 2013) chamam a atenção da área para a possibilidade de desenvolver pesquisas relacionados às novas transformações, incluindo o próprio fenômeno *Big Data*, Redes Sociais, *Cloud Computing* e *Web*

Semântica que aborda diretamente o objeto de estudo da área e que são possíveis causadores de impactos e mudanças paradigmáticas.

Pimenta (2013) destacava que apesar do *Big Data* ser o resultado da explosão informacional vivido atualmente, este também é a solução para o desafio da recuperação da informação em ambiente digital, um dos principais problemas da Ciência da Informação. Corrobora com a questão os autores Souza, Almeida e Baracho (2013) e Dias e Vieira (2013) de que o *Big Data* é um desafio para a Ciência da Informação no papel de conjugar as informações harmoniosamente. Dias e Vieira (2013) salientam ainda que a Ciência da Informação deveria prover de áreas de investigação que aportam conhecimentos associados ao novo tempo, à era da informação, na qual o *Big Data* é um dos muitos desafios.

Souza, Almeida e Baracho (2013) contribuem com o exposto ao apresentar que a Ciência da Informação tem um viés tanto social como aplicado e que, portanto, pode aproveitar as transformações para desenvolver teorias e se consolidar enquanto área do conhecimento com autonomia. Os autores também relacionam o impacto do *Big Data* em técnicas tradicionais de análise de assunto, análise documentária ou análise temática, pois a nova explosão informacional rompeu com a possibilidade do tratamento individual e intelectual dos registros. Além disso, exortam o profissional da Ciência da Informação a navegar nos espaços teóricos, adaptar-se aos contextos tecnológicos e reinventar-se continuamente (SOUZA; ALMEIDA; BARACHO, 2013).

No ano de 2014, às publicações nacionais referentes ao tema sinalizam a preocupação dos cientistas da informação para o desenvolvimento de estudos relacionados às novas transformações, visto que estas têm causado impactos em vários âmbitos e exigido novas abordagens para os estudos em Ciência da Informação e que implicam no futuro da área (BARACHO et a., 2014; RIBEIRO, 2014).

Em 2015 a área começa a discutir questões como a segurança da informação no *Big Data*. Os autores Rodrigues e Dias (2015) fazem uma análise aplicada dos impactos no vazamento de informação na *Web* e suas implicações nos modelos informacionais a partir do caso *Wikileaks*. Nesta perspectiva, Milagre e Santarém Segundo (2015) consideram que a Ciência da Informação pode aplicar as suas técnicas de organização do conhecimento, classificação e indexação ao *Big Data*, auxiliando nos processos e medidas de segurança da informação.

Neste mesmo período alguns autores abordam as implicações sociais e políticas do *Big Data* como desafios para a Ciência da Informação. Mostafa, Cruz e Amorim (2015) argumentam sobre o problema da teorização do *Big Data*, mais do que a análise do fenômeno em si. Segundo os autores, o *Big Data* é uma forma de capitalismo que transforma as pessoas em amostras, dados, mercados ou bancos. O *Big Data* mostra como os metadados medem as relações sociais, aprimoram o design do conhecimento "maquínico" e como monitoram e prevêem comportamentos de massas. Os referidos autores sinalizam ainda que é fundamental que áreas do conhecimento estudem para desvendar o *Big Data*, corroborando com a visão de Ribeiro (2014) e Santana Junior et al (2014) sobre a necessidade de se construírem teorias e desenvolver estudos relacionados para além das tecnologias. Baracho et al (2015) corroboram com a questão ao apresentar uma reflexão teórica de que a Ciência da Informação está se realinhando com suas origens para conectá-las às modernas tendências como o *Big Data*, redes sociais, *Web* semântica, entre outras transformações.

Prosseguindo no desenvolvimento de estudos sobre o *Big Data* na Ciência da Informação, Santana Junior (2016) discute o ciclo de vida dos dados, contemplando as fases de coleta, armazenamento, recuperação e descarte, na qual está presente transversal a este as fases de privacidade, integração, qualidade, direito autoral, disseminação e preservação. Para o autor supracitado, a Ciência da Informação pode oferecer um novo enfoque centrado nos dados, propondo soluções para o acesso, o uso e a manutenção dos mesmos, para assim garantir a qualidade, a recuperação e o descarte, assim como para gerar novo conhecimento a partir de dados não discriminatórios. Conforme apontam Coneglian, Santarém Segundo e Santana (2016), o *Big Data* está sendo conduzido a gerar resultados discriminatórios, durante o processo de análise de dados feito por algoritmos. Como alternativa, tem-se a necessidade de reflexões profundas dos dados que apresentam estes resultados e que a Ciência da Informação retrata tais questões, podendo debater os dados e entendê-los para garantir a qualidade e a análise dos mesmos.

Continuando na perspectiva de estudo da qualidade dos dados, Fagundes, Macedo e Freund (2017) e Furlan e Laurindo (2017) assistem que a Ciência da Informação pode contribuir com a aplicação de conceitos sobre dados e informações que vão além das tecnologias. Os referidos autores desenvolveram em seus estudos

um levantamento bibliométrico realizado na base de dados *Web of Science*, na qual fizeram uma identificação real das pesquisas realizadas sobre o *Big Data*, mostrando a importância que os cientistas da informação têm deferido ao fenômeno e as transformações emergentes.

Victorino et al (2017) discutem a proposta de construção de um ecossistema de *Big Data*, baseado nos aportes da Ciência da Informação como metadados, tesauros, taxonomias e ontologias para organizar e representar o enorme volume de dados, prestando suporte à análise de dados abertos governamentais conectados. Os autores discutem a qualidade e o ciclo de vida dos dados e esclarecem que a base do ecossistema de *Big Data* está em princípios e técnicas da Ciência da Informação que envolvem atividades de coleta, representação, armazenamento e disseminação seletiva da informação.

Barbosa e Kobashi (2017) evidenciam como a visualização de dados auxilia na busca e recuperação da informação utilizando de técnica da Ciência da Informação. Por seu turno, os autores Coneglian, Gonçalvez e Santarem Segundo (2017) exploram a intersecção entre as funções executadas por gestores, cientistas da computação e cientistas da informação em ambiente de *Big Data*, visando compreender o papel do cientista da informação neste novo contexto. De modo complementar, sinalizam que estes profissionais podem ser incluídos no processo de análise de dados fazendo uso dos aportes teóricos da própria Ciência da Informação.

Nota-se com as pesquisas aqui apresentadas uma evolução dos estudos de *Big Data* na Ciência da Informação. Na Figura 2, observamos essa evolução das temáticas abordadas pelos pesquisadores. Nos anos de 2013 a 2014, às pesquisas publicadas instigam a CI a prover de áreas de investigação que contemplassem o *Big Data*, as Redes Sociais, *Cloud Computing* e etc. Os autores que publicaram neste período apontavam e discorriam em seus textos que a área precisava expandir suas investigações e considerar aspectos do *Big Data*. Em seguida a área discute sobre a segurança da informação e as implicações do vazamento de informação na *Web.* Com a crescente discussão sobre o *Big Data* na área, os autores encetam pesquisas a respeito da qualidade e análise dos dados, considerando que o bom e correto gerenciamento, criação, manutenção e compartilhamento de dados podem auxiliar nos critérios mínimos de qualidade e ajudar a reduzir a incidência de resultados discriminatórios. Desta forma, os pesquisadores discutem o *Big Data* e suas

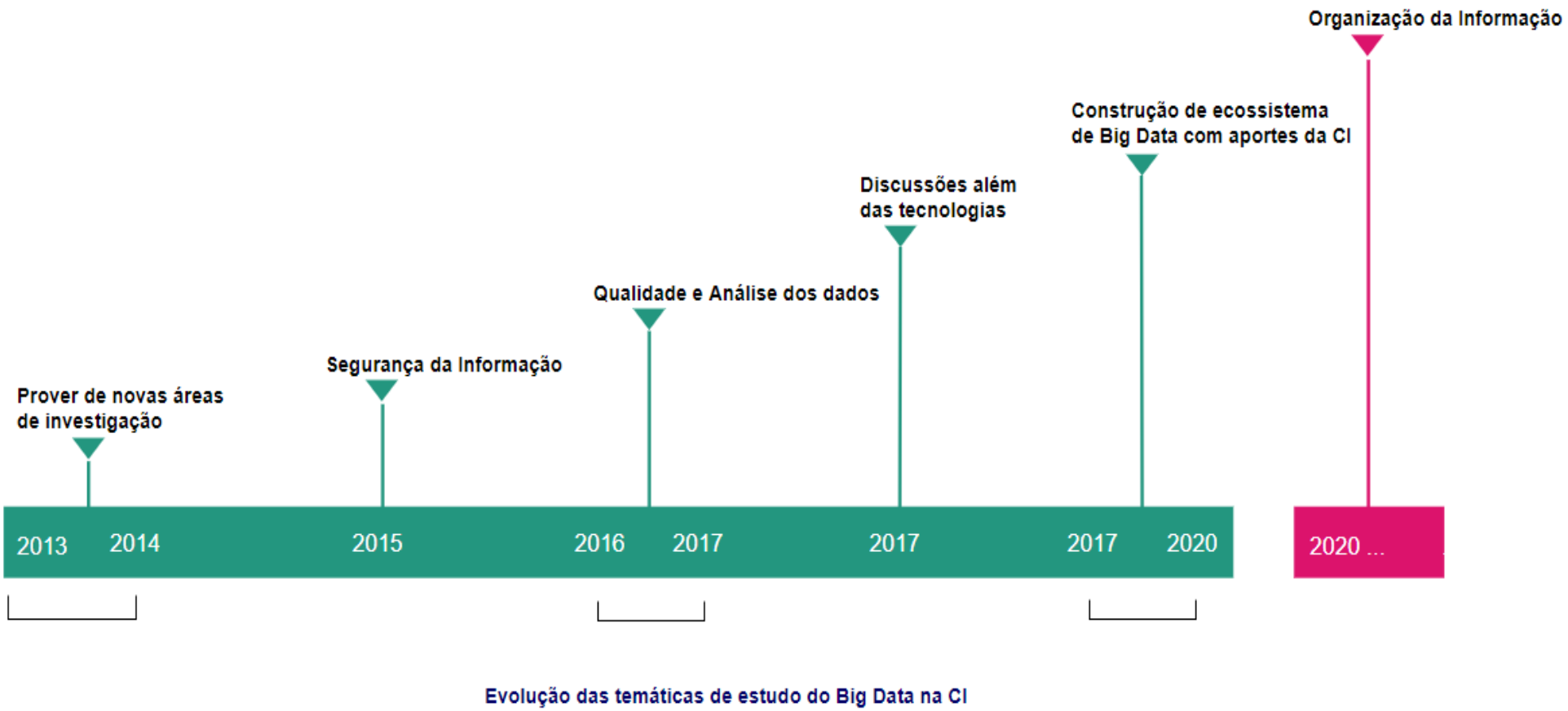
implicações para além dos aspectos tecnológicos, argumentando a necessidade de investigá-lo nos aspectos sociais, políticos, científicos, econômicos, culturais e outros. Discutem também o *Big Data* levando em consideração os aspectos da análise e da mitologia. No sentido das evoluções das pesquisas, a área propõem a partir de 2017 a construção de ecossistemas de *Big Data* com aportes para a mesma, ou seja, a criação de produtos, metodologias e ferramentas a partir das teorias, princípios, técnicas, conceitos e cânones da CI para serem usados nas diversas dimensões do *Big Data*.

A contar deste mesmo período, a área começa a observar e discutir que o aspecto mais importante é a organização da informação, ou seja, como organizar, gerenciar e compartilhar informações úteis, relevantes e confiáveis no contexto de transformações vigentes da sociedade do século XXI. Como conjugar as informações de forma harmoniosa para que se possa atender o usuário. Então as questões agora passam a vigorar em torno da organização da informação, como organizar, tratar, indexar e recuperar a informação no contexto do *Big Data*. Neste sentido, vigoram discussões sobre os impactos e mudanças paradigmáticas na CI, o papel do profissional da informação neste contexto, a análise dos dados, a interseção entre funções executadas por gestores, cientistas da computação e da informação em ambiente de *Big Data*.

Por conseguinte, a área que iniciou seus estudos com a teorização, discussões dos impactos e mudanças paradigmáticas do fenômeno em seu campo de conhecimento, objeto de estudo e profissional, perpassou por vários segmentos. Apesar de a área discutir o impacto do *Big Data* nos primeiros estudos, o fator tecnológico era o mais considerado. Todavia, com o avanço dos estudos nos segmentos de segurança da informação, ciclo de vida dos dados, qualidade dos dados, e etc, um fator mostrou-se mais emblemático, a organização da informação. Desta forma, a literatura nacional revela que o fator tecnológico não era mais a principal questão a ser discutida, entrando em cena questões relativas à organização de grandes volumes de dados e informações de forma útil, de fácil recuperação e que pudesse gerar conhecimento. Um desafio para as áreas do conhecimento e principalmente para a Ciência da Informação caracterizada como uma ciência dinâmica, dada à fluidez e ubiquidade de seu objeto de pesquisa.

Nesta seção identificamos as perspectivas de estudos sobre o *Big Data* na CI brasileira, ou seja, como a área começou a se articular no processo de discussão do tema. Na seção a seguir, discutimos as implicações do *Big Data* na Organização do Conhecimento, nos *Knowledge Organization System* (KOS), na classificação e indexação, de forma mais pontual.

Figura 2. Perspectivas de estudo sobre o *Big Data* na Ciência da Informação



Fonte: Autor

2.2 Implicações do *Big Data* na Organização do Conhecimento

As transformações tecnológicas, as mudanças de abordagem dos sujeitos informacionais com os sistemas advindas com a crescente autonomia dos usuários de criar, compartilhar e usar conteúdo na *Web*, entre outros fatores, têm gerado uma massa de volume informacional maior do que se pode medir e organizar individualmente. Em períodos anteriores à estas transformações, a sociedade já se referenciava ao campo de conhecimento da Ciência da Informação para obter técnicas e procedimentos que auxiliasse na organização para a recuperação da informação. Os *Knowledge Organization System* (KOS) surgem para apoiar as pessoas a encontrar informações relevantes e úteis, na qual o conhecimento é organizado em classes, conceitos, proposições, modelos, teorias e leis (HJORLAND, 2013).

Os profissionais dos KOs utilizam-se dos processos e instrumentos de classificação, catalogação e indexação para organizar os recursos informacionais. Porém, essas operações tradicionais da Organização do Conhecimento não conseguem abarcar o contexto das transformações vivenciadas atualmente. Os KOs e as teorias que os sustentam foram criados em contextos diferentes, com necessidades informacionais diferentes, com recursos tecnológicos diferentes e que para atender as necessidades do *Big Data* precisam ser atualizados. Este fenômeno rompeu com a possibilidade do tratamento individual e intelectual dos recursos, com a ideia de armazenamento físico do recurso, com a indexação feita exclusivamente pelo bibliotecário ou cientista da informação (HJORLAND, 2012, 2013; SOUZA; ALMEIDA; BARACHO, 2013; IBEKWE-SANJUAN; BOWKER, 2017).

No contexto do *Big Data* e da *Web* 2.0 os recursos informacionais são muitos, estão dispostos em vários locais, podem ser acessados por várias pessoas simultaneamente a qualquer momento e disponíveis, na maior parte das vezes, de forma completo. Enquanto isso, os KOs foram criados para organizar e recuperar recursos em suporte físico, com número limitado para o usuário e por unidade de informação, com necessidade da presença física para acesso e uso, sendo as atividades de classificação e indexação realizadas exclusivamente por profissionais, de forma unidirecional.

Frente ao exposto, os KOs tradicionais não podem ser utilizados da mesma forma para atender ao *Big Data*, pois o volume alto de dados dificulta estas práticas

em espaços como a *Web*. Idealmente, várias tecnologias de interação e produção do conhecimento encurtam os ciclos, reinventado os suportes físicos e tornando cada vez mais orgânica a relação com os registros de informação. Brooks (2001) menciona que o modelo clássico de representação do conhecimento pode não ser mais aplicável ou econômico no ambiente *Web* por se tratar de fenômeno heterogêneo, descentralizado e com alta taxa de crescimento. A solução proposta então é que os usuários personalizem os seus dados, representando de forma colaborativa os diversos recursos informacionais, digitalizados ou nato-digitais, no contexto *Web* (BROOKS, 2001).

Esta visão vai ao encontro com a proposta de Lancaster (2004), em que para certos tipos de materiais, a indexação orientada pelo o usuário pode até ser mais importante do que o é no caso de artigos de periódicos, livros, ou relatórios técnicos que têm seus metadados construídos de forma tradicional. Sendo a produção e a disseminação da informação cada vez mais distribuídas em diversas fontes, o conteúdo é variado e as visões de mundo também.

No entanto, no *Big Data* os sistemas tradicionais de organização do conhecimento não conseguem tratar o volume alto de informações que são produzidos de forma rápida e ininterrupta e em formatos variados, para organizar e recuperar informações relevantes e úteis. Segundo Ibekwe-Sanjuan e Bowker (2017) e Hjorland (2012), o *Big Data* e o Google não removeram a necessidade de classificação, apenas acentuam a necessidade da classificação. Cientistas e empresas que estudam o *Big Data* estão investindo mais na classificação, e a classificação construída de forma colaborativa entre milhares de usuários dispostos no ambiente digital.

Nesta ótica, destaca-se a classificação feita por seres humanos, visto que não é um algoritmo que determina os critérios de classificação, isso só pode ser feito por seres humanos (HJORLAND, 2012). Na classificação automática pode ser difícil obter um nível de qualidade suficiente. Além disso, o *Big Data* só funciona se for fornecido com bons metadados, tendo a coexistência de duas modalidades aparentemente opostas, o trabalho humano e a automação (IBEKWE-SANJUAN; BOWKER, 2017; HJORLAND, 2012).

A classificação humana no contexto do *Big Data*, conforme evidenciado por IbekweSanjuan e Bowker (2017) e Hjorland (2012), não remove a necessidade de

KOs construídos por profissionais, no entanto as transformações que o *Big Data* e a *Web* 2.0 proporcionam pressionam a área de Organização do Conhecimento a repensar o ponto de vista e a natureza que os KOs são projetados. Ibekwe-Sanjuan e Bowker (2017) relatam que a Organização do Conhecimento justificou a inclusão de termos e suas relações com base na garantia literária (tradicionalmente), e que no contexto do *Big Data* os KOs não podem se basear exclusivamente a elas, considerando-se as particularidades desta nova ambiência. Neste sentido, os KOs devem se adaptar a natureza mutável do seu produto, precisam se preocupar com as teorias, bem como adaptarem-se ao novo mundo de conhecimento, criarem esquemas maximamente flexíveis e repensarem a natureza do seu campo de atuação (HJORLAND, 2012; IBEKWE-SANJUAN E BOWKER, 2017).

Visto que o *Big Data* só funciona com bons metadados, os KOs precisam se valer de classificações facetadas, ao invés da aristotélica, de abordagens *top-down* a fim de atingirem o equilíbrio certo entre a classificação construída por especialistas e a classificação colaborativa (HJORLAND, 2012; IBEKWE-SANJUAN; BOWKER, 2017). Com a adoção destas atitudes, o número de dados relacionados ao recurso informacional pode ser significativamente alto, com variedade de termos relacionados aos aspectos físicos e intelectuais do recurso, os dados poderão refletir as visões de mundo dos usuários e as técnicas dos profissionais na representação do conhecimento.

Os KOs podem, assim, alavancar a participação pública e integrar a representação colaborativa da informação em alguns dos seus artefatos, visto que esta permite atualizar cabeçalhos de assuntos, aumentar a eficácia dos esquemas de classificação e dos sistemas de recuperação da informação, aumentar o nível de exaustividade dos termos, dentre outros (HJORLAND, 2012; IBEKWESANJUAN E BOWKER, 2017). A Folksonomia propicia obter facetas diversas do mesmo recurso informacional e de pessoas diferentes de maneira simples com o vocabulário natural do usuário, ou seja, usando o vocabulário não controlado. Como resultado, há a possibilidade de garantia de cobertura de assunto unindo está às classificações feitas por especialista.

Os sistemas de classificação construídos por especialista e a classificação construída por usuários contêm critérios funcionais diferentes, porém ambas devem ser utilizadas em conjunto na classificação e indexação de um recurso informacional,

principalmente em tempos de *Big Data* (HJORLAND, 2013; IBEKWE-SANJUAN; BOWKER, 2017). Os autores Ibekwe-Sanjuan e Bowker (2017) consideram que os modelos colaborativos entre cientistas e amadores passaram a representar esforços para alavancar e inserir a Folksonomia em artefatos dos KOs. Os autores também enfatizam a necessidade de libertar a construção de tesauros dos grilhões de regras de normalização da ISO de cima para baixo e integrar os termos de vocabulários não controlados.

Hjorland (2013) e Ibekwe-Sanjuan e Bowker (2017) apontam que uma das transformações que o *Big Data* causou na Organização do Conhecimento foi tornar possível a obsolescência de seus esquemas universais de classificação bibliográfica, por exemplo, por terem sido construídos com perspectivas e necessidades informacionais diferentes da atualidade. Segundo os autores é compreensível não esperar que as classificações bibliográficas, assim como os demais instrumentos da Organização do Conhecimento para organizar a estrutura semântica do conhecimento, mudem constantemente a cada descoberta, pois seria impossível fazê-la em tempo real e esta ação prejudicaria o usuário final na recuperação da informação. Todavia, que podem alavancar técnicas para construir KOs na qual dependam do conteúdo gerado pelo usuário.

Portanto, os KOs são centrais para o esforço científico e melhor enfrentamento do fenômeno *Big Data* pela Ciência da Informação. No entanto, precisam se adaptar às novas temporalidades do desenvolvimento teórico e técnico ocasionado pelo *Big Data* nas ciências sociais e naturais (HJORLAND, 2012, 2013; IBEKWE-SANJUAN; BOWKER, 2017).

Por conseguinte, quando falamos de *Big Data*, das implicações deste na CI e na Organização do Conhecimento, observamos que um dos produtos do mesmo são os dados. Os dados estão inseridos e se configuram como objeto de estudo e discussão na CI. Dentre os tipos de dados, existe os dados científicos na qual a sua disponibilização e compartilhamento torna-se fundamental no desenvolvimento da ciência. Portanto, na próxima seção apresentamos os dados científicos, suas tipologias, classificações, critérios e políticas quanto à preservação dos dados.

2.3 Dados científicos

A cultura do *Big Data* e seu impacto tecnológico, cultural e social, assim como a apropriação e uso do mesmo na Ciência da Informação e áreas correlatas, reivindica uma abordagem inovadora para lidar com os dados científicos (RODRIGUES; DIAS, 2017). Nesta conjectura há vários tipos de dados, e os dados científicos geram novos formatos com impacto científico, tecnológico e informacional, trazendo novas dinâmicas e formas de compreensão dos dados.

Lynch (2008) advoga que os dados científicos podem ser considerados *Big Data* por diversas formas, a primeira pela maneira que este desafiou o estado da arte das áreas relacionadas à computação, rede e armazenamento de dados e, em segundo, pelos dados científicos terem um grande valor, serem de uma significância duradoura e pelos desafios descritivos que podem exigir contexto. Desta forma, "os dados científicos abertos são facilmente compartilhados, replicados e recombináveis, eles apresentam tremendas oportunidades de reutilização, acelerando as investigações já em andamento e aproveitando os investimentos anteriores em ciência" (LYNCH, 2008, p. 28, tradução nossa).

A disponibilização e o compartilhamento dos dados científicos não é uma obrigação, mas uma necessidade frente aos desafios complexos que pressupõem a interdisciplinaridade, as colaborações entre laboratórios de pesquisa, permuta de informações e de competências que a sociedade enfrenta (AVENTURIER; ALENCAR, 2016). Quando o acesso aos dados científicos é limitado há uma sabotagem no processo de fazer ciência e, segundo os princípios para dados científicos descritos pela *Panton Principles*[5] não devem existir barreiras legais, financeiras ou técnicas neste acesso e as publicações científicas deveriam ser disponibilizadas em domínio público.

Os dados científicos são dados primários de uma pesquisa, referem-se a materiais não necessariamente textuais, incluindo produtos e/ou componentes que são disponibilizados abertamente sob licenças (ALBAGLI; CLINIO; RAYCHTOCK, 2014). O Relatório da *Organisation for Economic Co-operation and Development* (OECD)[6] descreve a expressão "dados de pesquisa" como "registros factuais usados

[5] Disponível em: http://pantonprinciples.org.
[6] OECD - Organisation for Economic Co-operation and Development. Disponível em: http://www.oecd.org/.

como fonte primária para a pesquisa científica e que são comumente aceitos pelos pesquisadores como necessários para validar os resultados do trabalho científico". Conforme a *National Science Board* (NSF)[7], essa definição inclui dados analisados e os metadados que descrevem como esses dados foram gerados.

Segundo Sayão e Sales (2016, p. 93) o termo dado científico "[...] tem uma amplitude de significados que vão se transformando de acordo com domínios científicos específicos, objetos de pesquisas, metodologias de geração e coleta de dados e muitas outras variáveis". Outrossim, podem ter várias aplicações e usos, além daqueles previstos pelos pesquisadores que o geraram.

Conforme a categorização proposta pela *National Science Board*, os dados científicos podem ser discriminados por sua natureza ou origem: observacionais, obtidos por meio de observação; computacionais, resultantes de execução de modelos computacionais ou de simulação; experimentais, provenientes de situações controladas em bancadas de laboratórios.

Por sua vez, a Universidade de Melbourne (2020, online) definiu os dados científicos como sendo:

> [..] fatos, observações ou experiências nas quais um argumento, teoria ou teste se baseia. Os dados podem ser numéricos, descritivos ou visuais. Os dados podem ser brutos ou analisados, experimentais ou observacionais. Os dados incluem: cadernos de laboratório; cadernos de campo; dados primários de pesquisa (incluindo dados de pesquisa em cópia impressa ou em formato legível por computador); questionários; fitas de áudio; fitas de vídeo; modelos; fotografias; filmes; respostas de teste. As coleções de pesquisa podem incluir slides; artefatos; espécimes; amostras. As informações de proveniência sobre os dados também podem ser incluídas: como, quando, onde foram coletadas e com o que (por exemplo, instrumento).

Partindo desta perspectiva Sayão e Sales (2016, p. 94) sustentam que, dependendo do ponto de vista, "[..] quase tudo que é gerado e coletado no ambiente de pesquisa pode ser considerado dados [científicos]". Os autores afirmam que os dados científicos não possuem valor sem a respectiva documentação que descreva seu contexto e as ferramentas utilizadas para criá-los, armazená-los, adaptá-los e analisá-los. Isso significa que a disponibilização de dados científicos na *Web* sem a

[7] Disponível em: https://www.nsf.gov/nsb/

devida contextualização impossibilita a sua interpretação e reuso, ao mesmo tempo em que inviabiliza a transmissão do conhecimento por ele aportado e reduz seu valor para a pesquisa interdisciplinar (SAYÃO, SALES, 2016).

Clinio e Albagli (2017) apresentam os dados científicos como uma inovação, não no sentido de melhoria incremental aos já conceituados formatos de divulgação da ciência como os artigos científicos. Mas como uma nova tecnologia literária que abre a totalidade dos registros de pesquisa para promover a produção coletiva e colaborativa de conhecimento, ampliar a participação na ciência, melhorar a qualidade da informação circulante e reestruturar o processo de avaliação por pares. Para Bradley (2012), artigos científicos relatam de maneira altamente condensada a elaboração de um experimento, oferecendo descrições genéricas insuficientes para sua replicação.

Nesta ótica, compreende-se que a disponibilidade dos dados científicos aumentaria muito as possibilidades de escrutínio, correção, refutação, complementação, colaboração, validação e aprendizado por um público amplo, diferente dos artigos científicos (CLINIO; ALBAGLI, 2017). Desta forma, estes dados se configurariam em um modelo de comunicação que adota práticas abertas de curadoria para explicitar os procedimentos que geraram os dados, permitindo ao pesquisador avaliar a relevância das afirmações e a qualidade dos dados acessando sua fonte (BRADLEY, 2010). Sob essa perspectiva, a ciência se basearia em evidências obtidas através do compartilhamento dos dados científicos, qualquer que seja seu *status* (em andamento, finalizado, descartado) ou resultado (parcial ou final; favorável ou ambíguo).

Os dados científicos estão presentes em todas as áreas e disciplinas do conhecimento, portanto estes podem variar em relação às abordagens dos diferentes atores (pesquisadores, instituições, financiadores) e em relação aos diversos contextos nacionais e internacionais nos quais estejam inseridos, podendo ser produzidos ou recolhidos de diferentes formas. Segundo Willis, Greenberg e White (2012, p. 1506, tradução nossa) devido a essa variedade e com o objetivo de “[...] ser útil para entender as semelhanças e diferenças, bem como o uso pretendido e potencial dos dados ao longo do tempo”, eles podem ser classificados ou agrupados de várias maneiras, seja pela disciplina que cria e planeja usar os dados ou pelo método de coleta empregado. Apesar disso, as diversas classificações e tipologias

existentes para os dados científicos implicam em escolhas que devem ser feitas para o arquivamento e a preservação (NSB, 2005).

Para uma melhor exposição dos enunciados discutidos neste ponto, o Quadro 5 apresenta a tipologia dos dados científicos proposta pela *National Science Foundation* (NSF)[8], conforme segue:

Quadro 5. Tipologia dos dados científicos

CRITÉRIOS	TIPOS	DESCRIÇÃO
Procedimento de coleta	Observacionais	Procedentes de pesquisa científica; caráter único, pois não se pode voltar a reproduzi-los; capturados em tempo real e geralmente fora de laboratório; registros de fatos ou evidências de fenômenos.
	Computacionais	Produtos da execução de modelos computacionais, simulações ou fluxos de trabalho; são reproduzíveis se preservadas a documentação de hardware e software, os dados de entrada e os passos intermediários.
	Experimentais	Procedentes de experimentos, procedimentos realizados em condições controladas com o fim de provar ou estabelecer hipóteses; caso seja um experimento replicável, os dados são mais fáceis de reutilizar e preservar.
Caráter	Primário	Trabalhos originais de pesquisa e/ou dados brutos sem interpretação coletados a partir de experimentos, pesquisas, entrevistas e demais técnicas; coletados a partir do problema de pesquisa; custo de obtenção maior; geralmente utilizados nas ciências, principalmente nas experimentais.
	Secundário	Menos precisos pois não foram coletados para responder ao problema de pesquisa; dados coletados, tabulados, ordenados e disponíveis publicamente em livros, periódicos, censos, biografias, artigos, bases de dados etc.; mais utilizados nas ciências sociais; permitem a repetição dos estudos e a criação de grandes conjuntos de dados, mais ricos e sofisticados.
	Terciário	Forma de dados derivados como recontagem, categorias e resultados de dados estatísticos; frequentemente utilizados para garantir a confidencialidade dos dados primários e secundários.
Grau de estruturação	Estruturados	Facilmente transferidos a outros sistemas devido à organização segundo um modelo definido; informação armazenada em tabelas e bases de dados relacionais seguindo uma estrutura determinada onde se definem as tabelas, os campos das tabelas e as relações entre ambos.
	Semiestruturados	Tradicionalmente inclui imagens, documentos de textos e outros objetos que fazem parte de uma base de dados; não tem um modelo de dados ou uma estrutura pré definida, não sendo possível manter em uma estrutura de base relacional; são irregulares e flexíveis, frequentemente adicionados hierarquicamente; possuem um conjunto consistente de

[8] A National Science Foundation (NSF) é uma agência federal independente criada pelo Congresso do Estados Unidos em 1950 "para promover o progresso da ciência; para promover a saúde, prosperidade e bem-estar nacional; para garantir a defesa nacional".

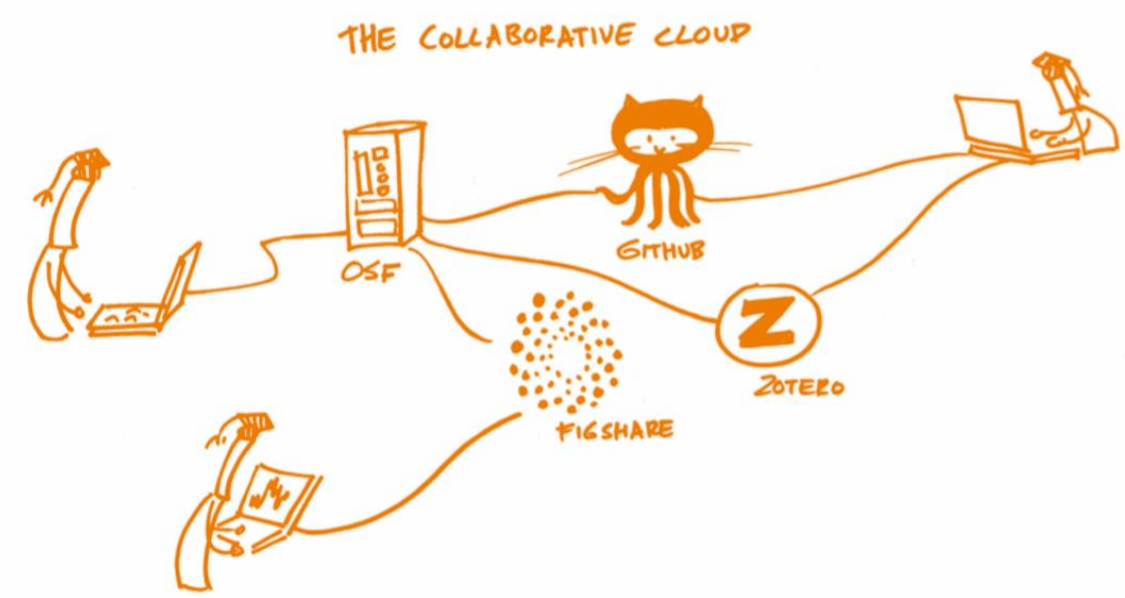
		conteúdo semântico e algum meio de classificá-los e ordená-los.
	Não estruturados	Não possuem um modelo definido ou estrutura identificável; cada elemento individual pode ter uma estrutura ou formato; nem todos os dados dentro de um conjunto possuem a mesma estrutura.
Nível de abertura	Dados abertos	Podem ser utilizados e distribuídos por qualquer pessoa, sem barreiras técnicas ou legais sob os requisitos de reconhecer a autoria e compartilhar o novo produto nas mesmas condições.
Formato	Arquivos digitais	Formato e o software com que se criam os dados dependem da forma que os pesquisadores coletam e analisam e geralmente são influenciados pelas normas e costumes de cada disciplina; grande quantidade de formatos disponíveis; opção mais segura para garantir o acesso; e o uso de formatos padrões.

Fonte: National Science Foundation (2020)

Cada um desses critérios e diferenças implicam em decisões políticas importantes em relação à preservação dos dados. A criação dos dados científicos acontece de maneira diferente em cada área do conhecimento, disciplina e subdisciplina, tendo o seu próprio conjunto de características de dados, tamanhos, tipos e níveis de complexidade. A compreensão destes dados vai depender do contexto em que eles são criados, portanto é necessário que os pesquisadores forneçam as informações sobre tal, mediante os metadados ou outros instrumentos.

Observa-se que segundo Semeler (2017, p. 64), “os dados científicos são descentralizados e gerados em diversos laboratórios e/ou centros de pesquisa por muitos pesquisadores em diferentes países do mundo”, conforme exemplificado na Figura 3. Nesse contexto, os dados científicos devem ser “válidos, compartilháveis, heterogêneos e contextualizáveis dentro de uma comunidade científica” (SEMELER, 2017, p. 65).

Figura 3. Rede de colaboração dos dados científicos



Fonte: Melero (2018)

Rauen (2018) destaca os principais benefícios percebidos a partir da disponibilização dos dados científicos, quais sejam: a) ampliar a colaboração entre pesquisadores da mesma área; b) aumentar a visibilidade da pesquisa; c) promover eficiência do investimento público em atividades científicas; d) promover a verificação dos resultados pela comunidade científica; e) democratizar e tornar transparente os resultados das pesquisas; e f) acelerar o processo de inovação. Neste sentido, a importância de compartilhar os dados científicos resume-se a:

> Promover la innovación y la reutilización de los datos que potencialmente puedan tener nuevos usos. Facilitar la colaboración entre usuarios de datos, creadores de datos y reutilizadores. Maximizar la transparencia y la fiabilidad de los datos. Favorecer la reproducibilidad de los ensayos experimentales. Permitir la verificación de los resultados de investigación. Reducir costes al evitar la duplicación de datos. Aumentar el impacto y la visibilidad de la investigación. Promover los proyectos de investigación de los que provienen los datos y sus publicaciones. Generar un reconocimiento directo de los investigadores productores de datos, como ocurre con cualquier otro resultado de investigación (MELERO, 2018, p. 8)[9].

[9] Promover a inovação e a reutilização de dados que possam ter novos usos. Facilitar a colaboração entre usuários, criadores e reutilizadores de dados. Maximizar a transparência e a confiabilidade dos dados. Promover a reprodutibilidade de testes experimentais. Permitir a verificação dos resultados da pesquisa. Reduzir custos, evitando a duplicação de dados. Aumentar o impacto e a visibilidade da pesquisa. Promover os projetos de pesquisa dos quais os dados provêm e suas publicações. Gerar reconhecimento direto dos pesquisadores produtores de dados, como ocorre com qualquer outro resultado de pesquisa (MELERO, 2018, p. 8, tradução nossa).

A disponibilização e o compartilhamento dos dados científicos é uma prática do movimento da Ciência Aberta, balizados na iniciativa Budapeste de Acesso Aberto. Segundo Aventurier e Alencar (2016, p. 4) “no contexto da ciência aberta e do aumento do volume de dados, os dados de pesquisas abertos apresentam potencial estratégico em diferentes níveis de uma organização para se explorar e valorizar seus conteúdos [...]”. De modo complementar, Lynch (2008) ressalta que para ser ativado a disponibilização, o compartilhamento e o reuso dos dados científicos, estes devem ser preservados, ao passo que os efeitos da perda dos dados podem ser de âmbito econômico ou pessoal e as experiências precisam ser executadas novamente, em outros casos pode representar uma oportunidade perdida para sempre. Além disso, o referido autor aponta que as instituições de fomento à pesquisa e outros financiadores agora veem corretamente os dados científicos como ativos que estão subscrevendo e, portanto, buscam o melhor retorno para os seus investimentos. Idealmente, os dados científicos e as possibilidades intimamente ligadas à eles “[...] exigem que pesquisadores e instituições documentem e implementem planos de gerenciamento e compartilhamento de dados que abordam o ciclo de vida completo dos dados - incluindo o que acontece após o término de uma concessão” (LYNCH, 2008, p. 28, tradução nossa).

As universidades encontram-se com obrigações legais[10] e éticas de fornecerem um legado dos dados científicos oriundos de pesquisas científicas conduzidas por sua comunidade científica. Tal necessidade culminou na criação e disponibilização de repositórios de dados científicos por diversas universidades e institutos de pesquisa do mundo. Dessa forma, oferecem plataformas de acesso aberto que visam publicar, conectar e preservar os registros científicos provenientes das pesquisas científicas conduzidas por suas comunidades.

Partindo dos pontos apresentados até este momento, temos o *Big Data* e que um dos produtos deste são os dados, e desta forma um dos tipos de dados são os dados científicos. Neste sentido há a necessidade de prover de uma infraestrutura que

[10] No contexto da América Latina e do Hemisfério Sul, temos ações como o documento vivo “Declaração do Panamá sobre Ciência Aberta”, a iniciativa *Open and Collaborative Science in Development Network* (OCSDNet), a publicação do “Manifesto pela Ciência Cidadã” em 2012. No Brasil temos a publicação do “Manifesto de Acesso Aberto a Dados da Pesquisa Brasileira para Ciência Cidadã” em 2016, o “4º Plano de Ação Nacional em Governo Aberto” publicado em 2018 com vigência até setembro de 2020, entre outros.

forneça garantia de confiabilidade e qualidade para possibilitar o gerenciamento, a disponibilização e o compartilhamento dos dados científicos, tendo em vista alcançar dois dos Vs aqui discutidos: Veracidade e Valor. Essa infraestrutura são os repositórios de dados científicos, apresentados na próxima seção.

2.3.1 Repositórios de dados científicos

Com a crescente visualização das necessidades e dos benefícios advindos do compartilhamento dos dados científicos, governos, universidades e instituições de fomento à atividade científica passaram a enveredar esforços visando a promoção do acesso aberto aos dados científicos. Por decorrência, tais instituições passaram a solicitar de seus beneficiários a adequação às práticas de compartilhamento de dados, desde a obrigatoriedade de um plano de gestão de dados científicos até as diretrizes para o compartilhamento em repositórios de dados (RAUEN, 2018). Segundo Sayão e Sales (2016, p. 93), nesse contexto de transição dois problemas são colocados:

> Por um lado os pesquisadores necessitam de infraestruturas que assegurem o máximo de confiabilidade, estabilidade e acessibilidade e que facilitem o trabalho de arquivamento, compartilhamento e reconhecimento de autoria para os seus dados; por outro lado, esses mesmos pesquisadores precisam encontrar coleções de dados de pesquisa, saber como acessá-las e sob que condições podem reutilizar esses dados e assim dar prosseguimento às suas pesquisas confiando na autenticidade e proveniência dos dados coletados ou gerados por outros pesquisadores.

Desta forma, essas instituições passaram a oferecer suporte aos pesquisadores quanto ao gerenciamento e compartilhamento dos dados científicos, por meio da construção de políticas, adequação e criação de infraestrutura tecnológica e de pessoal. Algumas instituições no desenvolvimento de suas estratégias expandiram seus serviços de repositórios digitais e/ou institucionais para receberem o depósito de dados científicos. Todavia, a literatura especializada no assunto revela que a melhor opção é o depósito dos dados em repositórios voltados, especificamente, para dados, “[...] sejam institucionais, disciplinares, multidisciplinares, orientados para projetos ou os ligados aos periódicos científicos” (SAYÃO; SALES, 2017, p. 68).

No centro deste arcabouço estão os repositórios de dados científicos que, além de oferecer uma base tecnológica para a contextualização dos dados, rapidamente se tornam parte essencial da infraestrutura mundial de pesquisa (SAYÃO; SALES, 2016).

Os repositórios de dados (*data repositories*), também conhecidos como repositórios de dados de pesquisa (*research data repositories*) ou repositórios de dados científicos (*scientific data repositories*) ou ainda *data centers*, são considerados parte importante da infraestrutura tecnológica para o compartilhamento e reutilização de dados (CURTY, 2015). Para Pampel et al (2013), Curty e Aventurier (2017) os repositórios de dados são serviços *online* que podem ser institucionais, temáticos, multidisciplinares, ligados à comunidades disciplinares e/ou resultantes de projetos de pesquisa específicos ou não, comumente todos os repositórios de dados enfatizam o acesso e a preservação dos dados científicos.

Curty (2015, p. 13, tradução nossa) evidencia que os repositórios de dados são "responsáveis pelo armazenamento, organização, curadoria e preservação de dados de pesquisa para acesso de longo prazo". Na concepção de Monteiro (2017, p. 19-34), são uma evolução dos primeiros repositórios documentais disponíveis em diversas universidades a partir de infraestrutura apropriada, capaz de dar suporte aos pesquisadores no gerenciamento e na disponibilização de dados científicos com fins de reuso por parte da comunidade científica.

Os autores Curty e Aventurier (2017) relatam que os primeiros repositórios de dados surgiram em 1960 e que, além da preservação a longo prazo, acesso e potencial reuso dos dados, os repositórios atuais possuem funcionalidades adicionais de manipulação e visualização dos dados, relatórios de estatísticas e métricas de uso. O *Registry of Research Data Repositories*[11] (re3data.org) afirma que cada vez mais universidades e centros de pesquisa estão construindo repositórios de dados científicos, possibilitando o acesso permanente aos dados científicos em ambiente confiável.

O re3data.org é um diretório que registra repositórios de dados científicos de diferentes disciplinas, atuando como uma ferramenta de ciência aberta que oferece a pesquisadores uma visão geral dos repositórios internacionais existentes. Atualmente, a ferramenta indexa, internacionalmente, cerca de 1900 repositórios de dados

[11] Disponível em: <https://www.re3data.org/>. Acesso em: 15 jan. 2020

científicos desde 2012. O *Edinburgh DataShare*[12] é um repositório que disponibiliza conjunto de dados científicos produzidos pela *University of Edinburgh*. Este é um repositório de dados científico institucional com escopo multidisciplinar, pois aceita dados científicos de várias áreas do conhecimento e são provenientes de diferentes instituições de pesquisa. Outro repositório multidisciplinar é o FigShare[13], desenvolvido pela Digital Science e a *Macmillan Publishers Company*, empresa internacional com sede nos Estados Unidos e no Reino Unido, permite que pesquisadores publiquem todos os seus dados de forma citável, pesquisável e compartilhável. Por sua vez, o PANGAEA[14] – *Data Publisher for Earth & Environmental Science*, é um repositório de dados científicos disciplinar. O mesmo disponibiliza dados georreferenciados sobre o sistema terrestre e desenvolvido e mantido pela *Alfred Wegener Institute for Polar and Marine Research* (AWI) e MARUM – *Center for Marine Environmental Sciences the University of Bremen*, na Alemanha. De modo geral, repositórios disciplinares ou temáticos "[...] são extremamente variados e heterogêneos, refletindo a multiplicidade de disciplinas e a diversidades de dados gerados no contexto da pesquisa científica mundial" (SAYÃO; SALES, 2016, p. 102).

Estes e outros exemplos de repositórios de dados evidenciam o importante papel desempenhado por esses sistemas para o acesso a uma ampla gama e tipos de dados científicos provenientes de diversas áreas do conhecimento, sua preservação e compartilhamento. Frente a isso, um dos desafios da disponibilização dos dados científicos é a gestão e a exploração do volume de dados produzidos. Portanto, a correta gestão dos dados científicos irá torná-los recursos informacionais passíveis de acesso e uso, ou seja, é neste processo que os elementos de representação são atribuídos, os quais "resumem todas as ações necessárias para tornar os dados passíveis de serem descobertos, acessíveis e compreensíveis ao longo do tempo: organização, documentação, armazenamento, compartilhamento e arquivamento" (LEIDEN UNIVERSITY, 2015, p. 1). Tem-se, portanto, o repositório de dados científicos como a infraestrutura central e norteador neste processo.

[12] DATASHARE. Disponível em: <http://datashare.is.ed.ac.uk>. Acesso em: 15 jan. 2020.
[13] FIGSHARE. Disponível em: <http:// figshare.com>. Acesso em: 15 jan. 2020.
[14] PANGAEA. Disponível em: <http://www.pangaea.de>. Acesso em: 15 jan. 2020.

Nesta perspectiva, instituições criadoras, mantenedoras ou financiadoras dos repositórios de dados têm desenvolvido e implementado políticas de gerenciamento, preservação e compartilhamento de dados científicos, com o objetivo de melhorar a eficiência da pesquisa, a colaboração e a transparência (LYON, 2007; SHEARER, 2015). Essas políticas devem estabelecer diretrizes e orientações para o depósito de dados em repositório de dados de acesso aberto.

A exemplo das agências de fomento, Lyon (2007) advoga sobre a importância de políticas construídas e implementadas para o gerenciamento, preservação e compartilhamento de dados. Para o autor supracitado, essas instituições devem tornar suas políticas claras, consistentes, acessíveis e apropriadas ao público pretendido. Além disso, devem abranger questões relativas à gestão e curadoria de dados, publicação, compartilhamento e preservação de dados, assim como "[...] devem levar em conta a paisagem em constante mudança de data centers e arquivos de dados, repositórios institucionais e de assunto, softwares de rede social, como wikis e blogs, e outras oportunidades de publicação na web (LYON, 2007, p. 46, tradução nossa).

Essas políticas melhoram as práticas de gerenciamento de dados científicos, favorecem o compartilhamento e reutilização por terceiros, possibilitam a verificação dos resultados da pesquisa e, finalmente, promovem outras inovações. Para tal, um elemento importante no contexto das políticas são os planos de gestão de dados, "[...] pois obrigam os pesquisadores a pensarem sobre como eles irão administrar seus dados antes do projeto, um requisito essencial para boas práticas de gestão de dados" (SHEARER, 2015, p. 4, tradução nossa). De modo geral,

> Um plano de gerenciamento de dados de pesquisa (DMP) define procedimentos e responsabilidades locais para o gerenciamento de dados e registros de pesquisa de um grupo ou projeto de pesquisa. O DMP deve, pelo menos, garantir que os requisitos desta política sejam atendidos, mas também deve incluir detalhes de requisitos locais específicos (por exemplo, aqueles relacionados à aprovação ética, propriedade intelectual e atribuição). (UNIVERSIDADE DE Melbourne).

Segundo a declaração da Organisation for Economic Co-operation and Development (OCDE) de 2003, as políticas de gestão de dados possibilitam a abertura, a transparência, conformidade legal, responsabilidade formal, profissionalismo, proteção da propriedade intelectual, interoperabilidade, qualidade e segurança, eficiência e prestação de contas. As características de uma política devem

refletir objetivos e princípios específicos nos quais ela se baseia. De acordo com Shearer (2015) uma política baseada no princípio do compartilhamento provavelmente se concentrará nas principais práticas necessárias para fornecer acesso aos dados científicos, enquanto que uma política baseada na administração dos dados se concentrará nas funções e responsabilidades envolvidas no gerenciamento dos mesmos. O autor contribui com a questão ao apresentar os elementos mais presentes e necessários nas políticas de gerenciamento de dados científicos, conforme segue:

Quadro 6. Elementos de uma política para gestão de dados científicos

REQUISITOS DA POLÍTICA	
Padrões e qualidade dos dados	Os investigadores são obrigados a aderir aos padrões internacionais para permitir o acesso e a reutilização.
	Documentação de dados e metadados devem acompanhar os dados para que sejam entendidos por outras pessoas.
Acesso e compartilhamento de dados	Os investigadores são obrigados a disponibilizar dados para serem compartilhados (geralmente mediante a publicação dos resultados ou pouco depois, embora algumas agências permitam períodos de embargo).
	Requisitos para depósito de metadados em catálogo local ou nacional.
Retenção e preservação de dados	Os dados devem ser retidos por determinado período de tempo mínimo.
	Quando possível, os investigadores devem depositar seus dados em arquivo de longo prazo para garantir a preservação de seus dados.
Planos de gestão de dados	As propostas de pesquisa devem incluir Plano de Gestão de Dados.
DISPOSIÇÕES COMUNS ÀS POLÍTICAS	
Privacidade	Os direitos e privacidade dos indivíduos que participam da pesquisa devem ser protegidos em todos os momentos. Os dados disponibilizados para uso mais amplo devem estar livres de identificadores que permitam ligações com participantes individuais da pesquisa e variáveis que possam levar à revelação dedutiva da identidade dos indivíduos.
Conhecimento tradicional	No caso de conhecimento local e tradicional, os direitos dos detentores do conhecimento não serão comprometidos.
Dados de natureza sensível	Liberação de dados pode causar danos. Aspectos específicos dos dados podem precisar ser protegidos (por exemplo, locais de ninhos de aves ameaçadas ou locais sagrados).
Propriedade intelectual/Propriedade dos dados	Pode ser necessário, ocasionalmente, atrasar a publicação por curto período, a fim de permitir que os pedidos sejam redigidos.

OUTROS ASPECTOS	
Princípios	As políticas de dados aderem a um conjunto de princípios abrangentes que articulam seu valor.
Escopo/Cobertura da política	Descreve o escopo dos dados cobertos pela política.
Papéis e responsabilidades	A política identifica as várias partes responsáveis pela gestão de dados nos diferentes estágios do seu ciclo de vida.
Monitoramento e execução	Os meios pelos quais as políticas serão monitoradas ou aplicadas são descritos na política.

Fonte: Shearer (2015, p. 8-9, tradução nossa).

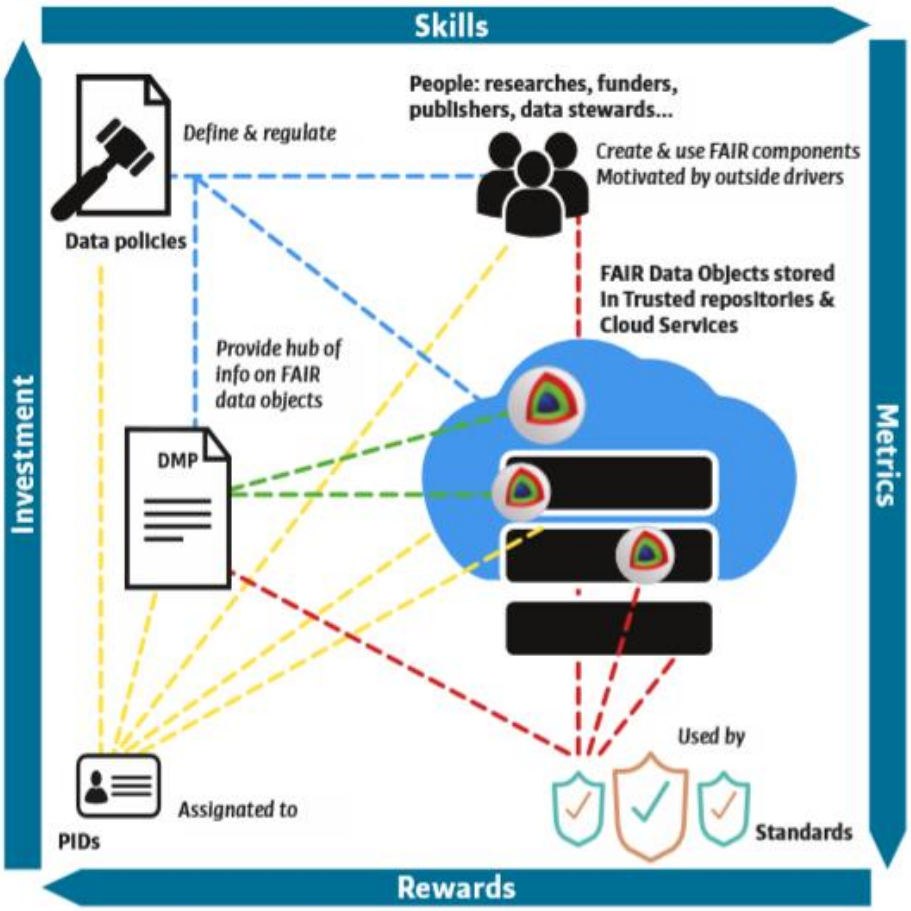
Como o exposto, as políticas de gerenciamento de dados científicos não se limitam a depositar os dados em uma base, mas também incluem informações sobre como e em qual contexto e motivação os mesmos foram gerados. Desta forma é fundamental esclarecer como os metadados serão armazenados, de modo a "[...] fornecer descrições sobre os conjuntos de dados, detalhando como eles foram produzidos, quando, onde e como podem ser reutilizados e também quem os gerou" (CAVALCANTI, 2018 apud PIERRO, 2018). No que se segue, a descrição correta dos metadados possibilita a padronização dos dados científicos, facilita seu acesso em repositórios de dados e o reuso em outras pesquisas com vistas a geração de conhecimento e inovações.

Logo, a qualidade do conjunto de dados científicos é um aspecto importante, sendo preciso definir os critérios para avaliar a qualidade desse conjunto de dados (AVENTURIER, 2017). Uma iniciativa neste sentido são os Princípios de Dados da FAIR (*Findable, Accessible, Interoperable* e *Reusable*), que apresentam uma diretriz para padronizar e aprimorar o gerenciamento de dados com quatro princípios fundamentais - Localização, Acessibilidade, Interoperabilidade e Reutilização. Segundo Aventurier (2017) o FAIR é um conjunto de princípios orientadores e práticas aceitas pela comunidade para que os produtores e os usuários, humanos ou não, possam usar mais facilmente os dados e citá-los corretamente.

Segundo Hodson et al. (2018), o ecossistema de dados FAIR é composto por: políticas que regulam e definem os dados; pesquisadores que os produzem ou os utilizam; planos; identificadores; padrões; metadados; repositórios confiáveis; e

serviços de nuvem em que os dados são armazenados. Por sua vez, esses componentes devem ser desenvolvidos dentro de uma estrutura proativa de quatro elementos principais: *skills, investment, metrics the rewards*, conforme descritos na Figura 4. Neste estudo, os autores apontam que os registros precisam ser desenvolvidos e implementados para todos os componentes e de tal maneira que eles saibam da existência um do outro e interajam. Os componentes de infraestrutura que são essenciais em contextos e campos específicos, devem ser claramente definidos. Os dados precisam ser representados em formatos padrões e ser acompanhados por identificadores únicos e persistentes (DOI), metadados e normas (HODSON et al. (2018).

Figura 4. Componentes de um ecossistema de dados



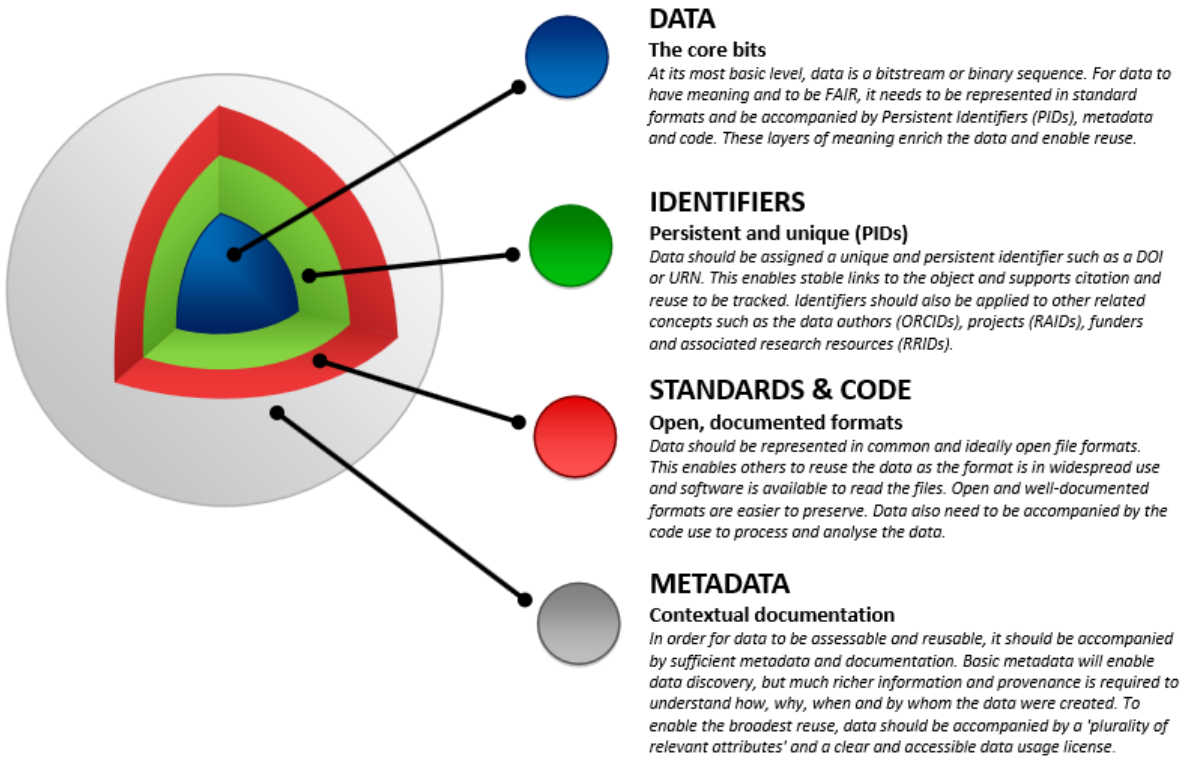
Fonte: Hodson et al., (2018, p. 6)

Os dados devem ser representados em formatos de arquivo comuns e idealmente abertos. Para que os dados sejam avaliáveis e reutilizáveis, eles devem

ser acompanhados por metadados e documentação. Para permitir a reutilização mais ampla, os dados devem ser acompanhados por uma “pluralidade de atributos relevantes” e uma licença de uso de dados clara e acessível (HODSON et al, 2018).

Conforme descrito na Figura 5 apresentada a seguir, o modelo de dados FAIR é composto por dados; identificadores persistentes; pela descrição de seus padrões e formatos utilizados na obtenção, representação, visualização dos dados, dentre outros; e metadados que descrevem os *datasets* que permitam a sua interpretação e reutilização. Desta forma os dados científicos serão localizáveis, acessíveis, interoperáveis e reutilizáveis.

Figura 5. Modelo de dados FAIR



Fonte: Hodson et al., (2018, p. 5)

Este ecossistema de informações e estratégias permite que os dados científicos sejam gerenciados, recuperados e preservados garantindo a qualidade dos dados e a confiabilidade dos sistemas de armazenamento. Sobre isso, Sayão e Sales (2016, p. 99) evidenciam que o repositório de dado, enquanto plataforma de acesso aberto “[...] integra diversas funções, tem como perspectiva oferecer um ambiente

dinâmico e flexível – principalmente pela natureza heterogênea dos dados - para dar apoio à execução dos processos de gestão de dados de pesquisa". Ademais, este sistema digital garante princípios de reprodutibilidade e de autocorreção da ciência, é uma infraestrutura sustentável e permanente de informações, apoia a validação e a revisão das publicações científicas, além de tornar parte da memória digital mais fidedigna da ciência. Entretanto, os repositórios de dados cumprem fidedignamente seus objetivos, se aliados com um ecossistema de políticas e estratégias de gerenciamento de dados científicos (SAYÃO; SALES, 2016).

No percurso desta pesquisa identificamos que os repositórios de dados científicos são apresentados como uma infraestrutura sustentável, confiável e viável para gerenciar um dos produtos do *Big Data*, os dados científicos. Recordando que quando discutimos sobre o impacto do *Big Data* na CI, os autores apontavam que a organização da informação é o cerne da questão há se solucionar no contexto do *Big Data*, ou seja, as discussões não precisam mais se restringir somente aos aspectos tecnológicos, mas considerar como organizar, tratar, indexar e recuperar as informações garantindo a relevância, a confiabilidade e o acesso ao público.

Desta forma, apresentamos uma infraestrutura possível e a partir desta circunstância há a necessidade de discutir a classificação e indexação destes dados científicos, tendo em vista que o *Big Data* tem um grande volume de dados, formatos variados e alta velocidade de produção dos mesmos. Os dados científicos se apresentam nesta infraestrutura em diversos formatos, podendo ser em imagem, texto, vídeo, gráficos, tabelas, dados estatísticos e muitos outros[15]. Sendo assim, nos questionamos como classificar, indexar e recuperar as informações nos repositórios de dados científicos, levando em consideração utilizar sistemas maximamente flexíveis.

Por conseguinte, apresentamos a representação colaborativa da informação (folksonomia) para a utilização nos ambientes digitais, mais especificamente nos repositórios de dados, como um sistema maximamente flexível. Observamos que esta permite a classificação e indexação de um recurso informacional, independente do formato, pelos usuários. Estes, por sua vez, podem indexar diversos recursos, com diversos termos, esses termos podem ser comum a outros recursos e a outras

[15] Ou seja, em formatos de *txt*, *pdf*, *png*, *jpeg, html, json, csv* entre muitos outros.

pessoas ao mesmo tempo. Desta forma, tem-se várias visões sobre o recurso informacional, prática que pode resultar em uma indexação mais exaustiva.

É observado que se pode estar trabalhando não apenas com a participação e a contribuição dos usuários amadores, mas também trabalhar com a colaboração do profissional no papel de fazer a curadoria dos termos e, ao mesmo tempo, envolver várias pessoas neste processo. Tendo em vista que a maioria dos repositórios de dados científicos são gerenciados por equipes formadas por profissionais da informação, como bibliotecários.

Neste contexto, na próxima seção apresentamos a representação colaborativa da informação em ambientes digitais, seus aspectos epistemológicos, conceituais e sistêmicos, além de incluir as abordagens teóricas desta na perspectiva do *Big Data*.

3 REPRESENTAÇÃO COLABORATIVA DA INFORMAÇÃO EM AMBIENTES DIGITAIS

A grande quantidade de informações, a variedade dos formatos e dos conteúdos de dados na *Web* cresce exponencialmente a cada dia, e deste modo novas formas de representar e recuperar as informações surgem a fim de otimizar estes processos. Segundo Barros (2011, p. 16), “no final dos anos 90 surgiram novas combinações tecnológicas que aumentaram a velocidade e a facilidade do uso de aplicações [no ambiente digital]”. Neste processo, a *Web* deixa de ser apenas publicadora de conteúdo e passa a ser interativa. Considerando o referido cenário, neste ponto da pesquisa são apresentadas as perspectivas da representação colaborativa da informação em ambientes digitais, tomando-se como base os aspectos epistemológicos, conceituais e sistêmicos, e apresentamos as abordagens teóricas desta na perspectiva do *Big Data*.

3.1 Aspectos epistemológicos

O envolvimento dos sujeitos em processos de organização da informação no ambiente digital tornou-se possível a partir da evolução da *Web* em *Web* 2.0 e suas possibilidades de interação e compartilhamento de informações. Neste momento “[...] uma quantidade sem precedentes de conteúdo começou a ser gerada por meio de registros da Web, wikis e outras ferramentas sociais, graças à redução de barreiras tecnológicas e de custos” (QUINTARELLI, 2005, p.1). Criadores de conteúdo surgem com a vontade de expressar suas opiniões e ideias e reconhecem na *Web* um espaço propício para a comunicação e divulgação de informações a partir do uso de vocabulários compartilhados para a classificação do conteúdo.

Emerge assim uma nova capacidade revolucionária de comunicação, colaboração e compartilhamento de informações *online* entre milhares de usuários dispostos na Internet, tornando-os ao mesmo tempo autores, editores, disseminadores e indexadores das informações (QUINTARELLI, 2005). Simultaneamente, ocasionado pela alta produção de documentos, Frederick Wilfrid Lancaster, um dos principais teóricos da representação temática, discorre sobre a indexação de recursos informacionais na Internet, em razão das dificuldades quanto à localização dos documentos devido a sua constante mudança, ausência de controle

de qualidade da informação e falta de coerência no uso dos termos (LANCASTER, 2004). Como alternativa, o autor apresenta a indexação de documentos pelos próprios usuários, visando dirimir os problemas de organização, representação e recuperação da informação no ambiente digital.

No contexto atual, sistemas populares de *software* social têm apresentado métodos que permitem aos usuários descrever e organizar o conteúdo com palavras-chave (*tag*) não estruturadas e explícitas, sem qualquer controle por parte do *Website*, que decorre na posterior recuperação da informação. Este tipo de indexação manual conduzida na *Web* é chamado de marcação, com termos de índice referidos como *tags* (VOSS, (2007). Na literatura especializada, diversas nomenclaturas são empregadas para descrever o resultado desta marcação colaborativa, podendo ser Folksonomia (QUINTARELLI, 2005), indexação social (HASSAN-MONTERO, 2006), indexação democrática (RAFFERTY; HIDDERLEY, 2007), etnoclassificação (MERHOLZ, 2004; BOYD, 2005; WALKER, 2005), classificação distribuída (MEJIAS, 2004), além de expressões mais populares entre pesquisadores e adeptos das ferramentas sociais, como etiquetagem social, etiquetagem colaborativa, classificação social (VOSS, 2007; FURNER & TÊNIS, 2006; LANDBECK, 2007; G. SMITH, 2004; TRANT, 2008) dentre outros.

Quintarelli (2005, p. 6) defende que “as Folksonomias não são uma teoria ou uma estratégia de cima para baixo: elas nasceram de um recurso (ferramentas de classificação popular) [...]”. Este sistema de organização foi popularizado no final da década de 90 e introduzido por *software* social como o Del.icio.us[16], Flickr[17], Technorati[18]i. Tratam-se de ferramentas para organizar páginas digitais que permitem aos usuários categorizar *sites* favoritos adicionando os próprios termos ou *tags*, além de compartilhar essa coleção com outras pessoas (SPITERI, 2007).

---

[16] Schachter (2004, *online*, tradução nossa) expõe a ferramenta Del.icio.us como sendo “um gerenciador de bookmarks sociais. Ele permite que você adicione facilmente sites que você gosta à sua coleção pessoal de links, classifique esses sites com palavras-chave e compartilhe sua coleção não apenas entre seus próprios navegadores e máquinas, mas também com outras pessoas”

[17] O Flickr, que se propõem ao gerenciamento e compartilhamento de fotos, é uma plataforma digital dedicada à hospedagem e partilha de imagens como fotografias, desenhos e ilustrações, além de permitir novas maneiras de organizar as fotos e vídeos. “Por seu alto nível de interatividade com os usuários, é um dos componentes mais exemplares da Web 2.0”. (WIKIPEDIA, *online*)

[18] A plataforma Technorati era dedicada à publicidade de editores que servia como uma solução de publicidade para os milhares de sites em sua rede. Enquanto junção das palavras tecnologia e literatos, o termo Technorati invoca a noção de inteligência tecnológica ou intelectualismo. (WIKIPEDIA, *online*).

Mathes (2004) em seus estudos observa que o diferencial destes *Websites* para outros existentes na época é a ênfase nas palavras-chave adicionadas pelo usuário como uma construção organizacional fundamental. Essas palavras-chave (*tags*) permitem que os usuários descrevam e organizem o conteúdo com qualquer vocabulário que escolherem, conforme a prática adotada pelo sistema Flickr, o qual possui sistema de marcação livre para fotos que foram adotadas e modeladas após o Del.icio.us. Nesta plataforma, as *tags* podem ser adicionadas no momento do *upload* ou quando as fotografias são exibidas pelo sistema.

Golder e Hubermam (2005) identificaram, a partir da condução de um estudo realizado na plataforma Del.icio.us, algumas funções das *tags* atribuídas pelos usuários, quais sejam: identificar os tópicos dos itens marcados; o tipo de coisa marcado (um blog ou um livro); o proprietário do conteúdo; categorias de refino (*tags* que refinam ou qualificam as categorias existentes); e auto referência (identificam o conteúdo em termos de sua relação com o autor da tag). Segundo Quintarelli (2005), a recompensa das pessoas ao usar a Folksonomia nestes *softwares sociais* era a visibilidade que suas *tags* adquiriam, gravitando em evidência no próprio sistema. Desta forma, a popularidade da marcação colaborativa na *Web* fez ressurgir o interesse pela indexação manual (VOSS, 2007).

Este sistema orgânico em desenvolvimento no Del.icio.us, Flickr e Technorati, que enfatizava o usuário como protagonista no processo de indexação manual de conteúdo na *Web*, a partir do vocabulário que lhe conviesse, foi intitulado de “Folksonomy” por Tomas Vander Wal numa lista de discussão sobre Arquitetura da Informação em 2004 (VANDER WAL, 2007).

3.2 Aspectos conceituais

Em âmbito mundial, a literatura sobre o assunto ainda está se desenvolvendo e caminhando para uma consolidação conceitual, epistemológica e sistêmica. O interesse por esta temática manifestou-se inicialmente através de artigos e discussões em *blogs* e, apenas posteriormente, iniciaram-se as primeiras publicações pelos meios científicos convencionais. Apesar de o primeiro sistema folksonômico a obter adesão massiva de usuários ter surgido em 2003 e do termo ter sido cunhado em

2004, as publicações científicas sobre o assunto surgiram em 2004 com os estudos de Mathes, Taylor, Merholz, e no Brasil o trabalho de Catarino e Baptista, publicado em 2007, seguido por Moura em 2009.

No mesmo período em que surge o termo Folksonomia, outros termos como "folkclassificação", "indexação colaborativa", "etiquetas colaborativas" e outros, já sinalizados nesta pesquisa, passam a ser empregados para designar a mesma ação. Porém, os conceitos atribuídos por Vander Wal ao termo Folksonomia é que se consolidam na literatura e no discurso de estudiosos na área, sendo justificada a opção pelo uso do termo Folksonomia na presente pesquisa para se referir a esta prática.

A terminologia cunhada por Thomas Vander Wal é um neologismo criado a partir da combinação de "*folk*" (povo, pessoas) e "*taxonomy*" (taxonomia), podendo-se ser descrita como sendo:

> Folksonomy is the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (usually shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information[19].

Wal (2007) considera que a Folksonomia pode ser descrita como uma estrutura categórica criada pelos usuários, de baixo para cima, desenvolvida com um tesauro emergente. Logo, os usuários criam as suas categorias de assunto baseadas nos seus próprios vocabulários. Trata-se de uma indexação livre e social em linguagem natural, em que não são adotadas regras, políticas ou instrumentos de indexação e nem o controle vocabular, sendo realizada pelos usuários do *software* social a qualquer momento.

Para Catarino e Baptista (2009) é necessário definir o que são informações ou objetos, que Wal define como "[...] qualquer coisa com uma URL". Diante disso, nesta pesquisa optou-se antes por utilizar o termo recurso, pois na definição do W3C o termo é empregado para se referir aos objetos (MILLER, 1998). Com efeito, a Folksonomia pode ser entendida como sendo "[...] o resultado da etiquetagem dos recursos da Web

[19] O resultado da marcação pessoal e livre de informações e objetos (qualquer coisa com uma URL) para a sua recuperação. A marcação é feita em um ambiente social (geralmente partilhada e aberta aos outros). A folksonomia é criada a partir do ato de marcação pela pessoa que consome a informação. (VANDER WAL, 2007, tradução nossa).

num ambiente social (compartilhado e aberto a outros) pelos próprios usuários visando a sua recuperação" (CATARINO; BAPTISTA, 2007, não paginado).

Baseado nas definições de Vander Wal, outros pesquisadores começam a estudar a Folksonomia e a se debruçar sobre o tema, tendo como resultado o surgimento de duas correntes de discussões quanto a sua definição:

- Folksonomia como resultado de um processo, ou seja, como um **produto** da atividade de etiquetagem do usuário; e
- Folksonomia como sistema, metodologia ou abordagem, ou o próprio **processo** em si.

A fim de melhor evidenciar a questão, o Quadro 7 contempla algumas definições de Folksonomia cunhadas por autores nacionais e internacionais que enfatizam as correntes de discussão ora descritas, a saber:

Quadro 7. Definições de Folksonomia (produto X processo)

PRODUTO	**DEFINIÇÃO**
Wal (2006)	Folksonomia é o resultado da atribuição livre e pessoal de etiquetas (*tagging*) a informações ou objetos (qualquer coisa com URL), visando à sua recuperação.
Lund et al. (2005)	Folksonomia se refere a um vocabulário, ou lista de termos, que surge da sobreposição de etiquetas definidas por vários usuários ao marcar as suas hiperligações favoritas, ou seja, seus marcadores para posterior recuperação.
Mathes (2004)	Folksonomia é um conjunto de termos que um grupo de usuários utilizou para etiquetar os conteúdos de recursos digitais da Web.
Trant (2006a, 2006b)	Folksonomia é o resultado de um sistema de classificação socialmente construído, ou, coleção de conceitos expressos num sistema de classificação desenvolvido de forma cooperativa. Folksonomia é um conjunto informal e orgânico de terminologia relacionada.
Sturtz (2006)	Folksonomia é um conjunto de etiquetas – com uma ou mais palavras-chave – que os usuários de um sistema compartilhado de gestão de conteúdo na Web aplicam a recursos individuais a fim de agrupá-los ou classificá-los para posterior recuperação
PROCESSO	**DEFINIÇÃO**

Russel (2005)	As folksonomias têm propiciado a possibilidade de criar desordenadamente, em texto livre, metadados atribuídos pelos usuários para recursos existentes (livros, imagens, URLs, etc).
Guy e Tonkin (2006)	Folksonomia é um tipo de sistema de classificação distribuída, a folksonomia é normalmente criada por um grupo de indivíduos, tipicamente os usuários do recurso.
Ohmukai, Hamasaki e Takeda (2006)	Trata-se de um sistema que administra etiquetas atribuídas pelos usuários aos recursos por eles indexados, compartilhando-as com outros usuários e também disponibilizando informações de outros recursos disponíveis na Web que foram indexados da mesma forma.
Quintarelli (2005)	Uma nova abordagem emergente para a classificação distribuída de recursos digitais.
Hammond et al. (2005)	Uma classificação não estruturada feita pelos próprios usuários dos recursos digitais.
Valongueiro (2006)	Um novo paradigma de classificação, pois respeita as diferenças culturais e características pessoais de quem utilizou e classificou determinada informação

Fonte: Catarino e Baptista (2009, p. 50)

De forma geral, estes autores interpretam a Folksonomia não só como um conjunto ou lista de termos que os usuários utilizam para indexar recursos informacionais no ambiente digital, mas como uma nova abordagem, metodologia, sistema de classificação e paradigma de classificação. Guedes (2010, p. 96) considera que a primeira linha de pensamento se justifica se "[...] pensarmos que uma taxonomia é o resultado de uma classificação terminológica de um determinado campo do conhecimento, assim, a Folksonomia seria o resultado funcional da classificação terminológica de um determinado usuário". Já na segunda linha de pensamento, a Folksonomia significaria todo o processo para se chegar ao resultado final. As duas linhas de pensamento não são excludentes, sendo ambas adotadas nesta pesquisa, tendo em vista as perspectivas que cada linha apresenta e pela pertinência que as mesmas contribuem para discorrer sobre aspectos de organização, representação e recuperação da informação digital em tempos de *Big Data*.

Outros termos relacionados ao conceito de Folksonomia enfatizam o aspecto colaborativo da ferramenta, com a palavra "social", a exemplo de social *bookmarking*, tal como descrito por Catarino e Baptista (2009), conforme segue.

Quadro 8. Termos relativos a indexação de recursos da *Web*

CONTEXTO	TERMOS	DEFINIÇÕES
Etiquetagem	Tagging	Tipo de ferramentas dá poder sem precedentes para os usuários que podem moldar as informações com as quais eles interagem (WINGET, 2006). É a atribuição de palavras-chave para classificar um objeto digital – fotografia, imagens, vídeos, áudio. Ou seja, é um processo de indexação de assuntos quase sempre sem um vocabulário controlado. (PATO, 2015, p. 87)
	Tagging Systems	Sistemas que habilitam usuários para acrescentar palavras-chave nos recursos digitais da Web sem o uso de vocabulários controlados (MARLOW et al., 2006)
	Social Tagging	Refere-se à prática de publicamente etiquetar ou categorizar recursos num ambiente compartilhado (TRANT, 2006b); ou um tipo de indexação aberta que se manifesta na Web (TENNIS, 2006)
	Social Tagging Systems	Permitem que os usuários compartilhem suas etiquetas de recursos particulares, além de que cada etiqueta serve como uma hiperligação para recursos adicionais que foram indexados por outros (MARLOW et al., 2006).
	Collaborative Tagging Systems	São sistemas colaborativos de etiquetagem que permitem, aos usuários, indexar as suas hiperligações, fotografias, referências e outros recursos digitais com palavras-chave ou etiquetas (VOSS, 2006); ou então processo pelo qual os usuários adicionam metadados em forma de palavras-chave ou etiquetas para compartilhar conteúdos (GOLDER; HUBERMAN, 2006a).
Classificação	Social Classification	Sinônimo de folksonomia que, para o autor, são metadados criados pelos próprios usuários da informação (SPITERI, 2007). Uma nova abordagem que está desafiando os esquemas tradicionais de classificação e de indexação baseados em vocabulários controlados (LIN et al., 2006). Processo pelo qual uma comunidade de usuários categoriza seus recursos naquela comunidade para o seu próprio uso (BOGERS; THOONE; BOSCH, 2006)
Bookmarking	Bookmarking	Um dos métodos mais populares para armazenar informação relevante da Web para acessá-la novamente e reutilizá-la (SPITERI, 2007).
	Social Bookmarking	Ferramentas que possibilitam que os usuários marquem suas páginas e atribuam etiquetas para representar seus temas de interesse (CAMPBELL, 2006)
	Social Bookmarking Manager	Denominação dada ao Delicious pelo seu criador. Golder e Huberman (2006b) definem este serviço como um sistema colaborativo para indexar os bookmarks da Web
Ontologia	Social Ontologies	Mote (2006) considera que o termo folksonomia *representa* social ontologies, ou seja, ontologias construídas de forma colaborativa, e significa uma classificação consensual gerada pelos usuários dos recursos digitais.
Taxonomia	Taxonomia Dinâmica	Joseph et al. (2006) afirmam que folksonomia é uma taxonomia dinâmica que representa as categorias que usuários individuais empregam para organizar seus espaços de informação

Fonte: Adaptado de Catarino e Baptista (2009, p. 52)

A partir dos conceitos e termos relativos à Folksonomia, observa-se que não há um consenso conceitual e terminológico a respeito da temática, tendo em vista que em alguns momentos ela é considerada produto, processo, ferramenta e/ou sistema. Além disso, é notável entre os autores o emprego de vários termos relacionados ao conceito de Folksonomia, sendo que alguns destes termos relacionam-se diretamente a ação de atribuir etiquetas (Etiquetagem e Classificação) e outros relacionam-se aos marcadores (*Bookmarking*). Não obstante, outros autores consideram o uso do termo inadequado, uma vez que o termo está erroneamente associado à taxonomia, sendo o termo mais apropriado "*ethnoclassification*" (MERHOLZ, 2004). Hammond et al (2005) já consideram que os termos “classificação social” ou “classificação distribuída” são mais adequados para representarem o fenômeno desta nova abordagem.

A partir do exposto, verifica-se a existência de várias abordagens, conceitos e termos relacionados às práticas folksonômicas, porém não há um consenso quanto a sua definição. Nota-se que as definições até então construídas consideram as principais características da Folksonomia, porém estas não são suficientes para representar este fenômeno em sua totalidade. Catarino e Baptista (2009) justificam esta situação a partir da tímida presença de literatura científica dedicada à construção da base teórica da Folksonomia, visto que o fenômeno tem um caráter muito mais pragmático do que teórico.

No entanto, mesmo com a falta de uma conceituação padronizada do significado do termo, fica evidente entre os pesquisadores da Ciência da Informação que a mesma está intimamente ligada à integração do usuário na representação da informação em ambiente digital, a partir do uso de *tags* para descrever conteúdos de recursos informacionais (MERHOLZ, 2004; MATHES, 2004; QUINTARELLI, 2005; HASSAN-MONTERO, 2006; VANDER WAL, 2007; BAPTISTA; CATARINO, 2007; SANTOS, CORRÊA, 2015a; IBEKWE-SANJUAN; BOWKER, 2017).

3.3 Aspectos sistêmicos

A Folksonomia retoma o uso da indexação manual, porém com uma perspectiva diferente. Nesta, os atores do processo de organização, representação e recuperação da informação são os próprios usuários dos ambientes digitais

colaborativos, possuindo padrões próprios em um espaço dinâmico, compartilhado e com atividades descentralizadas. Nesta sessão abordaremos as características, vantagens, desvantagens e algumas outras perspectivas relacionadas à Folksonomia.

Segundo os autores Hjorland (2013), Ibekwe-Sanjuan, Bowker (2017) e Massoni e Flores (2017) para que haja a existência de um sistema folksonômico três elementos e suas ligações são necessárias. Estes elementos são descritos por Silva e Silva (2008, p. 202) como sendo "[...] o usuário (*tagger*), o objeto e a *tag*. Uma folksonomia tem seu alicerce centrado na tag, que é o elemento de classificação para o objeto, dessa forma, uma atenção especial deve ser direcionada ao uso de termos (*tags*) em uma categorização." Guy e Tonkin (2006) indicam que essas *tags* podem ser palavras-chaves, categorias ou metadados, e podem ser classificados como qualquer palavra que define uma relação entre o recurso e o conceito na mente do usuário.

Na ótica de Hassan-Montero (2006), o processo do usuário atribuir uma *tag* ao recurso informacional envolve duas dimensões inter-relacionadas: a pessoal, e a coletiva e social. A primeira dimensão é o processo de indexação e categorização de recursos, cuja principal motivação é de caráter pessoal (egoísta), em que o usuário indexa o recurso para recuperar posteriormente, construindo um índice pessoal de *tags*. Neste caso, cada usuário confecciona seu próprio índice de *tags*. Por sua vez, a abordagem coletiva e social é a que confere maior potencial em áreas como a recuperação da informação, pois "quando os usuários compartilham as suas *tags* e recursos, geram mediante colaboração implícita um índice global de *tags* (Folksonomia) através do qual qualquer pessoa pode recuperar qualquer recurso descrito por outros usuários" (HASSAN-MONTERO, 2006, p. 1, tradução nossa). Essas duas características definem a Folksonomia como um novo modelo de indexação, em que os próprios usuários analisam o conteúdo de um documento e assimilam termos descritivos (*tags*), gerando um índice da qual podem recuperar vários outros.

Este novo modelo de indexação apresenta algumas características vantajosas como simplicidade no processo de representação do recurso, uso de termos não estruturados e hierarquizados, uso de termos flexíveis e navegação por entre etiquetas se comparado com as ontologias e taxonomias formais. Para Oliveira et al (2017) a Folksonomia ganha cada vez mais espaço por ser uma forma de classificação

que não utiliza taxonomias ou vocabulários pré-estabelecidos, não há restrição hierárquica pré-definida ou restrição para definir *tags*. Assim, algumas das características que a diferencia de outros sistemas incluem o próprio usuário é quem organiza a sua informação “mediante vocabulário próprio, viabiliza que o usuário compartilhe percepção de classificação com outros usuários, a classificação pode ser feita e alterada em qualquer momento, a classificação não é submetida a outros usuários e/ou profissionais”, troca de etiquetas menos utilizadas por outras mais utilizadas segundo Barros (2011, p. 20), além de que os usuários podem negociar significados com outros usuários.

Outras vantagens da Folksonomia, segundo Viana, Arakaki e Dal’Evedove (2019, p. 513) “[...] é que esta permite capturar metadados não-triviais e importantes, o *feedback* é imediato, se consolida em uma forma de comunicação assimétrica entre os usuários por meio dos metadados e reduz as barreiras à cooperação”. Idealmente, permite o manuseio de grande volume de dados. Catarino e Baptista (2009) destacam também que as *tags* são grafadas indistintamente em diferentes alfabetos e idiomas, na forma simples ou composta, singular ou plural. Portanto, a Folksonomia diminui o custo de categorização do conteúdo e a manutenção é perene e dinâmica; questões fundamentais quando se trata de volume exponencial de dados.

Mathes e Quintarelli (2004, 2005) relacionam o uso de vocabulário não controlado às principais limitações e fraquezas da Folksonomia. Assim como a ambiguidade das *tags*, podendo surgir à medida que os usuários aplicam a mesma *tag* de maneiras diferentes, tem-se a falta de controle de sinônimos e de siglas, aspectos frequentemente tratados de forma eficaz em vocabulários controlados. Apesar disso, Aquino (2007, p. 10) defende o uso da Folksonomia e discorre sobre a falta do uso de vocabulário controlado, indicando que

> Isso não quer dizer que o sistema seja uma desordem total [...]. Na verdade, trata-se de um mecanismo de representação, organização e recuperação de informações que não é feito por especialistas anônimos, o que muitas vezes pode limitar a busca por não trazer determinadas palavras-chaves, mas sim um modo onde os próprio indivíduos que buscam informação na rede ficam para representá-la, organizá-la e recuperá-la, realizando estas ações com base no senso comum e tendo assim um novo leque de opções ao efetuar uma pesquisa para encontrar algum dado.

Hassan-Montero (2006) enfatiza que, por o usuário realizar o etiquetamento com linguagem natural e não controlada e assinalando, na maioria das vezes, os termos que só têm significados para si mesmo, poderia haver *tags* com significados vazios para o coletivo, gerando conhecidos problemas como a polissemia e a sinonímia. No entanto, Quintarelli (2005) explica que nem todas as limitações são defeitos, tudo seria uma questão de escolha. Existe uma perda ao se utilizar da Folksonomia, mas os ganhos podem compensar, sobretudo quando se trata de gerenciamento e organização de dados e informações no ambiente digital. À vista disso, a Folksonomia não deve ser vista como um contraponto às outras modalidades de indexação, mas uma alternativa de organização e representação da informação no ambiente digital (GUEDES; DIAS, 2010).

Ao estudar os aspectos epistemológicos, conceituais e sistêmicos da representação colaborativa da informação, nota-se a necessidade de identificar quais as discussões relacionadas ao *Big Data* na literatura brasileira. Desta forma, apresentamos na próxima seção as abordagens teóricas das folksonomias na perspectiva do *Big Data* identificadas na literatura da Ciência da Informação brasileira.

3.4 Abordagens teóricas da representação colaborativa da informação na perspectiva do *Big Data*

As perspectivas teóricas na Ciência da Informação brasileira sobre a representação colaborativa da informação no contexto do *Big Data* são escassas e não acompanham as discussões internacionais conduzidas neste eixo investigativo (BROOKS, 2001; HJORLAND, 2012; IBEKWE-SANJUAN, BOWKER, 2017). Apesar da imprecisão conceitual e das interpretações em aberto na literatura especializada e nacional (SANTOS; CORRÊA, 2017), a popularização e ações de representação colaborativa da informação no ambiente digital segue a passos largos, viabilizadas pela interatividade no ciberespaço. A esse respeito, Viana, Dal'Evedove e Tartarotti (2019, p. 329) salientam que

> O contexto heterogêneo e interativo do ciberespaço alterou substancialmente os processos que envolvem a informação, desde a sua produção à reutilização e compartilhamento, sendo imprescindível a discussão dos elementos que implicam ou viabilizam o

> desenvolvimento de produtos e sistemas de informação adequados às demandas e práticas sociais

De modo geral, os estudos nacionais versam sobre a análise e/ou descrição das estratégias de representação colaborativa da informação utilizadas pelos usuários na etiquetagem de recursos em *sites* colaborativos (NASCIMENTO, 2008; RODRIGUES e MOREIRA, 2009, 2010, 2012; GALDO; VIERA e RODRIGUES, 2009; SANTOS, 2013; SANTINI e SOUZA, 2010; RODRIGUES, 2010; PEREIRA e CRUZ, 2010; NASCIMENTOS e NEVES, 2010; BARROS, 2011; SOUSA, 2012; ALVES; MOREIRA e MORAES, 2013; SOUSA e BENETTI, 2016; MASSONI e FLORES, 2017; SANTOS e CORRÊA, 2015a, 2015b; NASCIMENTO e CARVALHO, 2017). Nestes estudos, os autores enfatizam que o uso de das *tags* viabiliza identificar conceitos, lugares, períodos e outras características sem necessariamente conhecer e ver o recurso informacional. Além disso, buscam identificar os efeitos da folksonomia na organização e na recuperação da informação, preocupando-se em conhecer as estratégias, preferências, motivações e interesses dos usuários.

Em outro eixo investigativo, os estudos versam sobre as diferenças da Folksonomia em relação às linguagens documentárias, seus principais conceitos, características, vantagens e desvantagens. O entendimento dos perfis, características culturais e cognitivas dos usuários bem como a sua intervenção nas atividades de organização de recursos informacionais também são contemplados (CATARINO e BAPTISTA, 2007, 2009, 2010; BRANDT, 2009; BRANDT e MEDEIROS, 2010; CARVALHO; LUCAS e GONÇALVES, 2010; STREHL, 2011; VIERA e GARRIDO, 2011; SANTANA, 2013; VIGNOLI; ALMEIDA e CATARINO, 2014).

A Folksonomia na perspectiva semiótica e/ou enquanto manifestação de linguagens criadas e compartilhadas pelos usuários é um foco investigativo que ganhou destaque na Ciência da Informação brasileira entre o período de 2009 a 2013 e retornando em 2019, tendo como representantes os estudos conduzidos por Moura (2009), Guedes (2010), Assis (2011), Assis e Moura (2011 e 2013), Guedes, Moura e Dias (2011 e 2012) e Viana, Dal'Evedove e Gracioso (2019). Os estudos permitem obter a compreensão e dimensão dos desdobramentos da participação ativa dos usuários na construção de linguagem para a organização e recuperação da informação em ambientes colaborativos. A prática da Folksonomia em espaços

sociais semânticos evidencia o poder da linguagem como mecanismo de interação visando o alcance dos significados de informação nos processos de análise documentária. Neste foco investigativo o contexto não pode ser desconsiderado, “[...] pois anulá-lo enquanto componente influente da linguagem interfere diretamente na forma como os interlocutores recebem a informação transmitida no processo de comunicação.” (VIANA, DAL’EVEDOVE, GRACIOSO, 2019, p. 82).

As autoras Viana, Dal’Evedove e Gracioso (2019) discorre em sua pesquisa as dimensões pragmáticas do processo de Indexação social[20], tendo por base a pragmática de Wittgenstein (1994), e a Teoria dos Atos de Fala, de Austin (1990). Segundo as referidas autoras a pragmática de Wittgenstein e os atos de fala de Austin podem:

> [...] auxiliar, também, no enriquecimento da indexação social, contribuindo para uma representação da informação mais inclusiva, participativa e representativa e, sobretudo, favorecendo o aprofundamento de discussões na perspectiva da dimensão cultural da organização do conhecimento. (VIANA, DAL’EVEDOVE, GRACIOSO, 2019, p. 82)

A proposição de metodologias destinadas à hibridização ou coexistência dos vocabulários controlados e da Folksonomia para a representação da informação em ambientes digitais foi objeto de estudo nos trabalhos apresentados por Santarém (2010ª e 2010b); Santarém e Vidotti (2011), Silva (2013 e 2014), Santos e Corrêa (2015a) e Santos, (2016). Nestes estudos, a Folksonomia é investigada na percepção da atuação dos usuários, enquanto mentes tradutoras e propositoras de novos arranjos e categorização. Os autores, para diminuir as desvantagens da Folksonomia, apresentam a alternativa de controlar o nível de liberdade dos usuários ao atribuírem as *tags*. Neste âmbito os cientistas da informação são desafiados a repensarem a construção de sistemas e metodologias frente à representação da informação em ambientes digitais.

Os aspectos relativos à importância da Folksonomia como ferramenta auxiliar aos instrumentos de controle terminológico voltados para a representação da informação também são temas de estudo (MOURA, 2009; GRACIOSO, 2010; SANTOS e CORRÊA, 2015b; TARTAROTTI; DAL’EVEDOVE e FUJITA, 2015;

[20] Um dos termos sinônimos a Representação Colaborativa da Informação.

BRIGIDI e PEREIRA, 2016; OLIVEIRA et al, 2017; SANTOS et al, 2017). Neste eixo, as pesquisas enfatizam a construção colaborativa de políticas de indexação, catálogos bibliográficos *online* de bibliotecas universitárias e ontologias com abordagens da representação colaborativa da informação. Os autores enfatizam tanto a possibilidade de uma indexação híbrida aliando os termos provenientes dos vocabulários controlados das bibliotecas universitárias aos termos atribuídos por autores dos recursos informacionais localizados nas palavras-chaves, resumos e na ficha de identificação da obra. Assemelham-se, portanto, aos estudos que trabalham na perspectiva da Folksonomia assistida, como forma de controlar o nível de liberdade do usuário e garantir a qualidade no processo de representação e recuperação da informação e manter o caráter dinâmico do processo de etiquetagem em ambiente digital (MOURA, 2009; GRACIOSO, 2010; SANTOS e CORRÊA, 2015b; TARTAROTTI; DAL'EVEDOVE e FUJITA, 2015; BRIGIDI e PEREIRA, 2016; OLIVEIRA et al, 2017; SANTOS et al, 2017).

O uso da Folksonomia no processo de representação e organização da informação imagética gerada na Rede e também como fator de construção da memória coletiva tem como representantes Aquino (2008), Pato (2015), Oliveira e Vital (2015), Nobrega e Manini (2016) e Gonçalves e Assis (2016). Nesses estudos os autores propõem a Folksonomia como uma nova forma de tratamento de imagem, consideram-se que a partir da representação do conteúdo no espaço digital, a preservação da informação imagética e construção da memória coletiva são contemplados. Neste pensamento, a Folksonomia se apresenta como elemento potencializador da memória coletiva, por meio do hibridismo das manifestações dos usuários em ambientes digitais, construindo uma memória coletiva em rede.

Em trabalhos como o de Viana, Arakaki e Dal'Evedove (2019) a folksonomia é apresentada como uma tendência para a construção de metadados confiáveis e representativos. Segundo as autoras:

> A criação e desenvolvimento de metadados de alta qualidade, acurácia, atualidade, consistência e completude é um dos atuais esforços da Ciência da Informação. Como contributo à questão, métodos mais eficazes de construção de metadados podem ser implementados a partir da inserção de tags oriundas da representação e descrição colaborativa dos recursos informacionais por diferentes usuários. (VIANA, ARAKAKI, DAL'EVEDOVE; 2019, p. 520).

As produções científicas apresentadas acima mencionam de forma geral algumas características relacionadas ao *Big Data*, como variedade de fontes e formatos de dados, velocidade na produção e disseminação destes dados e o volume grande dos mesmos. Tais características são mencionadas de forma esporádica em alguns trabalhos para contextualizar e justificar a representação colaborativa da informação em ambiente digital, mas sem alusão direta com o *Big Data* e seus impactos, riscos, desafios, benefícios e implicações do fenômeno. As publicações de Aquino (2008) e Moura (2009), por exemplo, mencionam o problema do grande volume de dados, a velocidade com que as informações surgem e a dificuldade para recuperá-las e que as Folksonomias auxiliam neste processo, mas sem relacioná-los ao *Big Data*.

Dentre os estudos, alguns autores contextualizam as suas investigações diante das emergentes transformações ocorridas na sociedade, a exemplo do desenvolvimento tecnológico, da grande produção de informação e do papel das bibliotecas universitárias em prover acesso às informações de qualidade. Tartarotti, Dal'Evedove e Fujita (2015) apontam a importância da organização da informação no ambiente *Web*, frente a variedade dos formatos dos recursos informacionais existentes no ambiente digital e ressaltam que questões relacionadas com o conhecimento no mundo digital devem ser incluídas como objeto de pesquisa do campo científico da Organização do Conhecimento.

Brigidi e Pereira (2016), Santos e Corrêa (2015a) e Nascimento e Carvalho (2017) mencionam o volume de informações e a variedade dos formatos no ambiente digital como fatores que causam problemas na alimentação de dados nos sistemas de informação. Citam o fator tecnologia que tem gerado novas formas de interação, comunicação, produção, acesso, recuperação e apropriação em diferentes dispositivos e sistemas.

Viana, Arakaki e Dal'Evedove (2019) mencionam os diversos formatos e a maior necessidade de classificação que o *Big Data* ocasionou. Segundo as autoras a *World Wide Web* desencadeou uma maior produção e compartilhamento de recursos e objetos informacionais, tornando a representação destes um desafio híbrido, "[...] em que se busca conciliar o entendimento sob uma ótica subjetiva do ponto de vista humano, considerando a multiplicidade compreensiva dos termos, e objetiva para o processamento formal de máquinas" (VIANA, ARAKAKI, DAL'EVEDOVE; 2019, p.

510). As autoras ainda destacam que as ações colaborativas em ambientes digitais podem favorecer a criação, gerenciamento e manutenção de metadados mais representativos e com valor agregado.

As publicações aqui analisadas, mesmo que mencionam alguns aspectos e características que possam estar relacionados ao *Big Data*, não apresentam em seus objetivos e reflexões o mesmo como ponto de discussão relacionados à representação colaborativa da informação. Portanto após a análise destas publicações científicas, salvo os trabalhos de Viana, Arakaki, Dal'evedove (2019), Tartarotti, Dal'Evedove e Fujita (2015) e esta dissertação, identificamos a inexistência de pesquisas dedicadas à representação colaborativa da informação com intersecções diretas ao *Big Data*.

Na próxima seção desta dissertação, apresentamos como se configurou a representação de dados científicos na rede de repositórios da FAPESP e identificamos a existência, ou não, de práticas de representação colaborativa. A busca por estudar estas ações na rede se baliza por observar, a partir dos autores aqui citados, a necessidade de incluir sistemas flexíveis para a indexação, representação e recuperação dos dados científicos e que atuassem em consonância com as características do *Big Data*. Tendo em vista as sinalizações da literatura científica brasileira e estrangeira e práticas como as observadas na *LibraryThing*[21].

Também pela rede se apresentar como uma iniciativa pioneira, sintonizada com as práticas de *Open Science,* quanto a disponibilização dos dados associados às pesquisas desenvolvidas em todas as áreas do conhecimento no Estado de São Paulo. Principalmente, pela proposta da FAPESP de tornar a disponibilização dos dados científicos, por meio da rede, acessível para o público comum, ou seja, o público que não é pesquisador, que não é acadêmico, que não é cientista.

Desta forma, na próxima seção apresentamos os resultados desta pesquisa e discutimos sobre a FAPESP, a rede de repositórios, as instituições participantes, as categorias básicas, as políticas e os ambientes digitais.

[21] LibraryThing é um aplicativo de catalogação social para armazenar e compartilhar catálogos de livros e vários tipos de metadados de livros. Baseado em Portland, Maine, o LibraryThing foi desenvolvido por Tim Spalding e foi ao ar em 29 de agosto de 2005.

4 ANÁLISE DA REDE DE REPOSITÓRIOS DE DADOS CIENTÍFICOS DO ESTADO DE SÃO PAULO

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) é uma instituição de natureza pública[22] de fomento à pesquisa científica ligada à Secretaria de Desenvolvimento Econômico, Ciência, Tecnologia e Inovação do governo do estado de São Paulo. A instituição começou a ser implementada em 1960 no governo de Carvalho Pinto por Paulo Vanzolini e pela comissão da Universidade de São Paulo sob o reitor Ulhôa Cintra.

Seu principal objetivo é conceder apoio financeiro e manter o cadastro dos projetos de pesquisa desenvolvidos no Estado de São Paulo, custear a instalação de novas unidades de pesquisa e manter o cadastro das mesmas, fiscalizar a aplicação dos auxílios, promover estudos sobre o estado geral da pesquisa em São Paulo e no Brasil e promover ou subvencionar a publicação dos resultados das pesquisas[23]. Os seus programas tem mantido como princípios a valorização da prática da ciência básica como indissoluvelmente ligada à qualidade de ensino e à capacitação para inovações tecnológicas.

A FAPESP tem consolidado seu pioneirismo tanto no sistema de financiamento, pelas Unidades da Federação, à pesquisa quanto na disponibilização dos dados associados às pesquisas desenvolvidas em todas as áreas do conhecimento no Estado de São Paulo.

A iniciativa da FAPESP com a Rede de Repositórios de Dados Científicos das Universidades do Estado de São Paulo[24] foi forjada a partir de um movimento internacional de disponibilização de dados científicos, observados em agências de fomento públicas e privadas da América do Norte, Austrália, Grã-Bretanha, Holanda, Alemanha e países escandinavos.

Estas instituições de fomento por observância, outrora, as políticas de gestão de dados desenvolvidas por seus governantes, outrora, por garantir o maior benefício possível para o avanço científico e tecnológico, inicializaram a exigência de um plano de gerenciamento de dados como componente obrigatório na fase de submissão de

[22] Com dotação mínima do orçamento do Estado de São Paulo, não menor que 0,5%, destinados à pesquisa científica e tecnológica. Devendo garantir sua autonomia através de administração privativa de suas verbas. (HAMBURGER *et al*, 2004)

[23] Lei n. 5.918, de 18 de outubro de 1960. Disponível em: https://www.al.sp.gov.br/repositorio/legislacao/lei/1960/lei-5918-18.10.1960.html

[24] Rede de Repositórios de Dados Científicos das Universidades do Estado de São Paulo. Disponível em: https://metabuscador.uspdigital.usp.br/

projetos. Para as instituições o compartilhamento de dados reforça a investigação científica aberta, incentiva a diversidade de análises e opiniões, promove novas pesquisas, possibilita o teste de hipóteses e métodos de análise novos ou alternativos, facilita a formação de novos pesquisadores, possibilita a exploração de tópicos não previstos pelos investigadores iniciais e permite a criação de novos conjuntos de dados quando dados de várias fontes são combinados[25].

Nestas regiões o plano de gerenciamento de dados é um exemplo prático da maneira pela qual patrocinadores públicos e instituições de pesquisa estão implementando a Ciência Aberta, o esforço para tornar a pesquisa e os dados científicos livremente acessíveis.

Neste sentido, a rede de repositórios de dados científicos promovida pela FAPESP é uma iniciativa prática da gestão dos dados de pesquisa no Brasil sintonizada com as práticas de *Open Science*. Lançado em 16 de dezembro de 2019, a rede de repositórios reúne dados científicos de oito instituições do Estado de São Paulo, a saber:

- Embrapa Informática Agropecuária (CNPTIA/EMBRAPA)
- Instituto Tecnológico de Aeronáutica (ITA)
- Universidade de São Paulo (USP)
- Universidade Estadual de Campinas (UNICAMP)
- Universidade Estadual Paulista (UNESP)
- Universidade Federal do ABC (UFABC)
- Universidade Federal de São Carlos (UFSCar)
- Universidade Federal de São Paulo (UNIFESP)

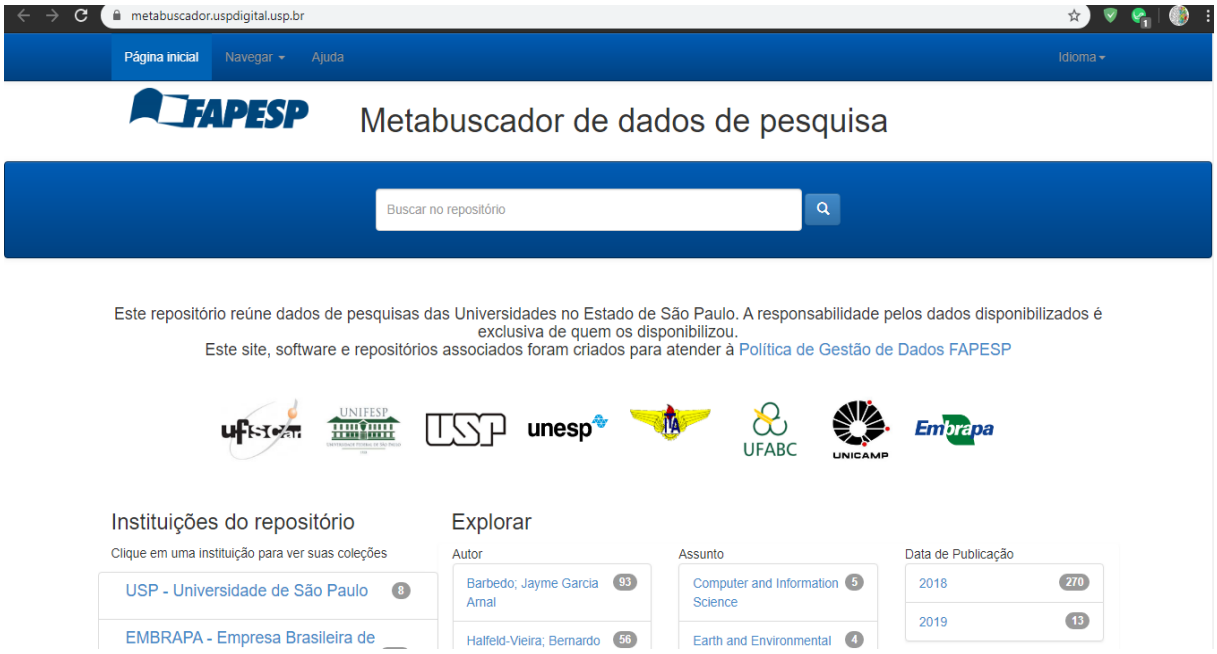
A colaboração entre as instituições participantes iniciou em 2017 a partir da exigência pela FAPESP, de um Plano de Gestão de Dados, para determinadas modalidades e chamadas, entre os anexos obrigatórios de propostas submetidas (MEDEIROS, 2019). Segundo a Agência FAPESP[26] (2019), a rede disponibiliza de modo organizado em uma plataforma aberta, dados associados às pesquisas desenvolvidas em todas as áreas do conhecimento das instituições mencionadas. Por

[25] GRANTS & FUNDING. Disponível em: https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#goals
[26] AGÊNCIA FAPESP. Disponível em: http://agencia.fapesp.br/fapesp-lanca-rede-de-repositorios-de-dados-cientificos-do-estado-de-sao-paulo/32251/

meio da plataforma é possível ter acesso aos dados gerados em pesquisas científicas, independentemente de sua publicação em artigos científicos. Para o pesquisador que gerou os dados, a Rede de Repositórios aumenta a visibilidade da sua pesquisa, permitindo o seu compartilhamento e reuso em pesquisas futuras.

Segundo Medeiros (2019), cada instituição integrante da rede desenvolveu seu próprio repositório de dados científicos e grupos permanentes internos para o gerenciamento e compartilhamento dos dados. A integração dos repositórios é viabilizada por um portal único, apresentado na Figura 6, que busca e disponibiliza informações de forma integrada. O portal de acesso, um buscador de metadados, foi desenvolvido pela USP utilizando o *software* DSpace[27] disponível nos idiomas português e inglês. O metabuscador permite navegar por comunidades e coleções (instituição), data do documento, autor, assunto, título e palavras-chave no campo de pesquisa. O usuário tem a opção de fazer a busca em todo o repositório ou especificar para o repositório desejado, além da opção de pesquisa avançada que permite a especificação nos campos que deseja pesquisar e combinar as pesquisas com os operadores booleanos "e", "ou" ou "não". Para acessar, visualizar e realizar *download* dos dados científicos dispostos na rede não é necessário que o usuário realize cadastro, todo e qualquer cidadão pode acessar, sem restrição.

Figura 6. Página inicial da Rede de Repositório FAPESP



[27] DSpace é um software de código-fonte aberto que fornece facilidades para o gerenciamento de acervo digital, muito utilizado para implementação de repositórios institucionais.

Fonte: USP digital

O *site*, o *software* e os repositórios associados foram criados para atender à Política de Gestão de Dados da FAPESP. A política aborda aspectos quanto à relevância do gerenciamento e compartilhamento de dados científicos para o avanço da ciência e da tecnologia, à racionalização de recursos, à facilidade da reprodutibilidade da pesquisa, além do treinamento de novos pesquisadores e exploração de aspectos não previstos no projeto original.

4.1 Apresentação e discussão dos resultados

A avaliação da Rede de Repositórios de Dados Científicos do Estado de São Paulo visando identificar as diretrizes e práticas adotadas para a organização e o tratamento de dados científicos pelo sistema versou em três eixos investigativos dedicados às categorias básicas, políticas da rede e ambiente digital. Consecutivamente, as mesmas discorrem sobre os aspectos necessários para a existência de um repositório, a forma com que os dados científicos são indexados no sistema e a atual estrutura do sistema em análise.

4.1.1 Categorias básicas

Neste ponto foram identificados os profissionais e as instituições responsáveis pela gestão da Rede de Repositórios de Dados Científicos do Estado de São Paulo e como os mesmos organizam e fazem a manutenção dos dados científicos. A instituição FAPESP foi a responsável por elaborar as políticas de gestão de dados e o plano de gestão de dados adotados pelos pesquisadores financiados pela agência de fomento. Como resultado, a Rede de Repositórios vai disponibilizar dados científicos de todas as áreas do conhecimento hospedados dos repositórios institucionais integrantes, construídos a partir das especificações tecnológicas e administrativas de cada uma das instituições participantes.

A FAPESP reuniu as instituições participantes e adotou algumas ações institucionais, como administrativas, técnicas (infra computacional), modelos de planos de gestão de dados e treinamento. As ações administrativas envolveram a criação de grupos de trabalhos permanentes representados por CNPTIA - Carla

Geovana do Nascimento Macario, ITA - Vera Lucia Junqueira, Emilia Villani, UFABC – Flavio Horita, Maria do Carmo Kersnowsky, UFSCAR – Ana Carolina Simionato, Denilson de Oliveira Sarvo, UNESP – Flávia Maria Bastos, Paulo Noronha Lisboa Filho, UNICAMP – Benilton Carvalho, UNIFESP – Maria Eduarda Puga, Flavio Costa de Souza, USP – Fatima Nunes Marques, Joao Eduardo Ferreira, METABUSCADOR – Diego Araujo, Edmar Martinelli (Superintendência de TI da USP-SC).

Cada instituição participante criou seus grupos de trabalhos e mecanismos para disponibilizar e tratar os dados científicos e orientar os pesquisadores sobre a produção e gerenciamento dos dados. Da mesma forma, cada instituição desenvolve e administra o seu próprio sistema, que é integrado em uma interface única de busca de metadados, comum a todas as instituições envolvidas, conforme exemplificado na Figura 7.

Figura 7. Interação do sistema da Rede FAPESP

Fonte: Adaptado de Medeiros (2019)

Esta interface, Metabuscador de dados de pesquisa, foi desenvolvido pela Superintendência de Tecnologia da Informação (STI) da USP e permite buscar por instituição, autor, assunto, ano ou palavras-chave. O *harvester* de metadados foi implementado em DSpace, a coleta de metadados é feita diariamente, aceita

comunicação com DSpace, Dataverse[28] e CKAN[29], usa OAI – PMH[30], Dublin Core qualificado e apresenta um conjunto de cerca de 200 campos de metadados.

Para a organização e manutenção dos dados, cada instituição desenvolveu suas próprias políticas e planos de gerenciamento de dados, sendo o DMP Tool[31] utilizado pela maioria das instituições. Os usuários podem fazer o *download* dos dados científicos independente de serem ou não cadastrados nos repositórios de dados. Além disso, tem como estratégia principal de captação de dados os dados científicos produzidos por pesquisadores financiados pela FAPESP. As instituições da Rede de Repositórios se responsabilizam pela preservação digital dos dados científicos disponibilizados pelos pesquisadores nos aspectos organizacionais, legais e técnicos.

4.1.1.1 Universidade de São Paulo

A USP designou por meio da portaria do reitor de 12 de junho de 2018, os membros do Grupo de Trabalho – Gestão e Repositório de Dados Científicos, com a incumbência de viabilizar uma infraestrutura computacional que possibilite o armazenamento dos planos de gestão de dados, bem como do repositório de dados científicos gerados pelos pesquisadores USP. A instituição em conjunto com seu grupo de trabalho, publicou e regulamentou a Resolução 7900, de 11 de dezembro de 2019, que estabelece normas para a gestão de dados científicos.

A resolução declara como elegível para o uso do repositório os docentes que tenham responsabilidade sobre a publicação dos dados científicos. Segundo a resolução, as condições para o armazenamento dos dados na plataforma são: o preenchimento dos metadados necessários de forma adequada e adequação aos aspectos éticos e legais quanto aos dados científicos. Para a utilização do repositório, uma solicitação deve ser realizada por meio de preenchimento de formulário

[28] Dataverse é um aplicativo da web de código aberto para compartilhar, preservar, citar, explorar e analisar dados de pesquisa.
[29] CKAN (*Comprehensive Knowledge Archive Network*) é um sistema livre de depósito e gerenciamento de dados que oferece ferramentas para publicação, compartilhamento, descoberta e uso de dados. É um sistema voltado a governos nacionais e regionais, companhias e organização que coletam muitos dados.
[30] OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting é um protocolo que define um mecanismo para coleta de registros de metadados em repositórios.
[31] DMP Tool é um serviço gratuito que ajuda pesquisadores e instituições a criar planos de gerenciamento de dados de alta qualidade que atendem aos requisitos dos financiadores, criado pelo Centro de Curadoria da Universidade da Califórnia (UC3). No caso, a instituição FAPESP está direcionando o uso da ferramenta para o servidor na Califórnia.

específico disponível no sistema corporativo do Repositório de Dados Científicos da USP.

Os pesquisadores da USP podem disponibilizar seus dados em qualquer formato e devem, da mesma forma, fornecer os metadados (descrição dos dados) a fim de facilitar sua compreensão e seu reuso para serem publicados na plataforma, ficando a responsabilidade de manter os mesmos seguros por um determinado tempo para a instituição. Após a solicitação do pesquisador, o Grupo Gestor de Dados Científicos, formado por integrantes da Pró-Reitoria de Pesquisa (PRP), Pró-Reitoria de Pós-Graduação (PRPG) e Agência USP de Gestão da Informação Acadêmica (AGUIA) realiza a análise.

Após a liberação pelo grupo gestor, far-se-á a inclusão dos dados no repositório, seguido de uma curadoria pela biblioteca e publicação no repositório USP. Assim como as regulamentações criadas pela USP, a mesma proveu seu *website* de informações e diretrizes a respeito da gestão e plano de dados científicos, sua organização e armazenamento, publicação, preservação digital e recomendações gerais. A Figura 8, apresenta a tela da interface do repositório de dados científicos da USP.

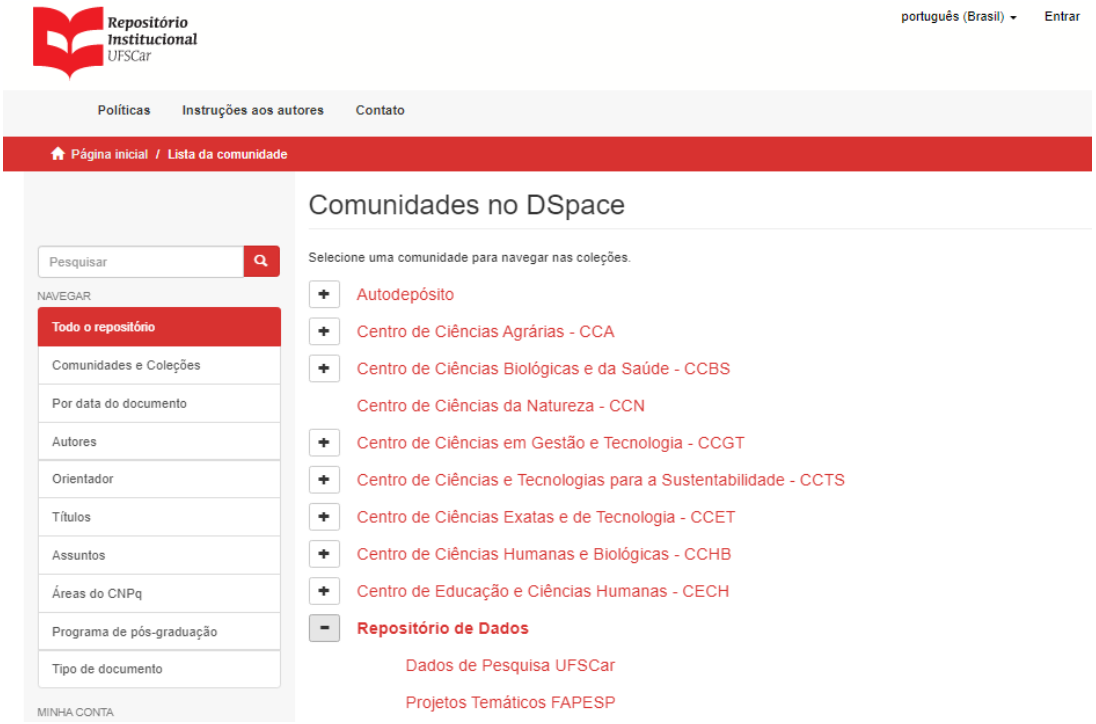
Figura 8. Repositório de Dados Científicos da USP

Fonte: USP (2020)

4.1.1.2 Universidade Federal de São Carlos

A UFSCar criou um projeto multidisciplinar, envolvendo a atuação de diferentes setores da universidade como a Secretaria de Informática (SIn), Pró-Reitoria de Pesquisa (ProPq), Departamento de Ciência da Informação (DCI), Sistema Integrado de Bibliotecas (SIBi) e Repositório Institucional (RI/UFSCar). O repositório de dados da instituição é integrado como uma coleção do Repositório Institucional da instituição (Figura 9), o qual contempla a coleção Dados de Pesquisa[32] e a coleção Projetos Temáticos FAPESP[33].

Figura 9. Coleção Repositório de Dados UFSCar



Fonte: UFSCar (2020)

Neste repositório podem ser depositados os dados científicos de docentes com vínculo na instituição, sendo adotado o autoarquivamento como método de

[32] Correspondem aos dados depositados pelos docentes da UFSCar provenientes de outras fontes de pesquisa.
[33] São dados de pesquisa vinculados aos projetos temáticos FAPESP de docentes da UFSCar.

povoamento com os próprios autores depositantes de seus dados. Com a restrição de depósito somente para docentes, é necessário a solicitação de permissão de uso para o primeiro depósito na coleção. Para auxiliar os docentes no processo do autoarquivamento, a instituição criou o documento administrativo Manual de Autodepósito de Dados de Pesquisa, disponível como consulta e *download* no sistema.

4.1.1.3 Universidade Federal do ABC

A UFABC criou seu grupo de trabalho no final do ano de 2017 com o intuito de elaborar políticas e diretrizes para os repositórios, identificando necessidades, tecnologias e orçamentos, além de promover a articulação com outras Instituições Federais de Ensino Superior (IFES) e implantar e disponibilizar as versões Beta dos repositórios. A instituição entrou para o programa de pesquisa Horizonte 2020, organizado pela União Europeia que investe em investigação e inovação, fazendo parte do eixo de trabalho “Dados Abertos de Pesquisa”.

A UFABC criou grupos de trabalhos diferentes, um para atuar no repositório institucional que utiliza o *software* Dspace, e outro para atuar com o repositório de dados científicos utilizando o *software* Dataverse. O grupo de trabalho foi composto por servidores do Sistema de Bibliotecas (SisBi), servidores do Núcleo de Tecnologia de Informação (NTI) e servidores da Pró-Reitoria de Pesquisa.

A instituição está desenvolvendo a política de uso do Repositório de Dados de Pesquisa-UFABC, que irá determinar o conteúdo, a sua estrutura informacional, as formas de depósito, as funções de cada integrante da equipe gestora e as questões legais do repositório. A instituição configurou e personalizou o Dataverse, tendo como padrões de organização, gerenciamento, compartilhamento e exploração de dados, assim como tipo de arquivos suportados e outros, prescritos pelo *software* em seu guia.

Como estratégia de captação de dados, a instituição instituiu a Portaria nº 451, de 22 de novembro de 2019, sobre o Plano de Dados Abertos 2020-2022 que estabelece responsabilidades e fluxos entre unidades administrativas a respeito da

coleta, organização e publicação de informações institucionais. A interface inicial do Repositório de Dados de Pesquisa da UFABC é apresentada abaixo.

Figura 10. Repositório de Dados de Pesquisa da UFABC

Fonte: UFABC (2020)

4.1.1.4 Universidade Federal de São Paulo

Em maio de 2017 a UNIFESP criou o seu grupo de estudos e trabalho para a implantação do repositório de dados de pesquisa (Figura 11). O grupo foi instituído pela Portaria da Reitoria n° 2036 de 02 de julho de 2018 e composto por membros da Reitora, da Assessoria de integração acadêmica, Pró-reitor adjunto de Pós-graduação e Pesquisa, diretoras da Coordenadoria da Rede de Bibliotecas da Unifesp (CRBU), divisão de tecnologia da informação, comitê gestor do repositório institucional, procuradora educacional Institucional e técnica em arquivos.

A instituição adota o *software* Dataverse e distribui a sua rede de dados de pesquisa em dois *datacenters*, um na cidade de São Paulo e outro localizado em

Diadema, no interior do Estado de São Paulo. A mesma sugere o uso do DMPonline[34] para a criação dos planos de gestão de dados e para prever antecipadamente o tamanho de cada depósito.

Figura 11. Repositório de Dados de Pesquisa da UNIFESP

Fonte: UNIFESP (2020)

Em relação ao gerenciamento dos dados, a UNIFESP criou o E-Dados[35] para ampliar o uso de dados e indicadores no planejamento estratégico da instituição e para apresentá-los à comunidade de forma geral. O E-Dados foi impulsionado pela constituição da Política de Gestão de Dados Estratégicos Institucionais da UNIFESP (Resolução 178, de 13 de novembro de 2019), que faz parte de uma abrangente Política de Dados da mesma, que também inclui a Política Nacional de Dados Abertos (e Plano de Dados Abertos da Unifesp 2018- 2019) e a Política de Dados de Pesquisa. Esta última está em processo de desenvolvimento e o escritório tem a responsabilidade de implantação e manutenção.

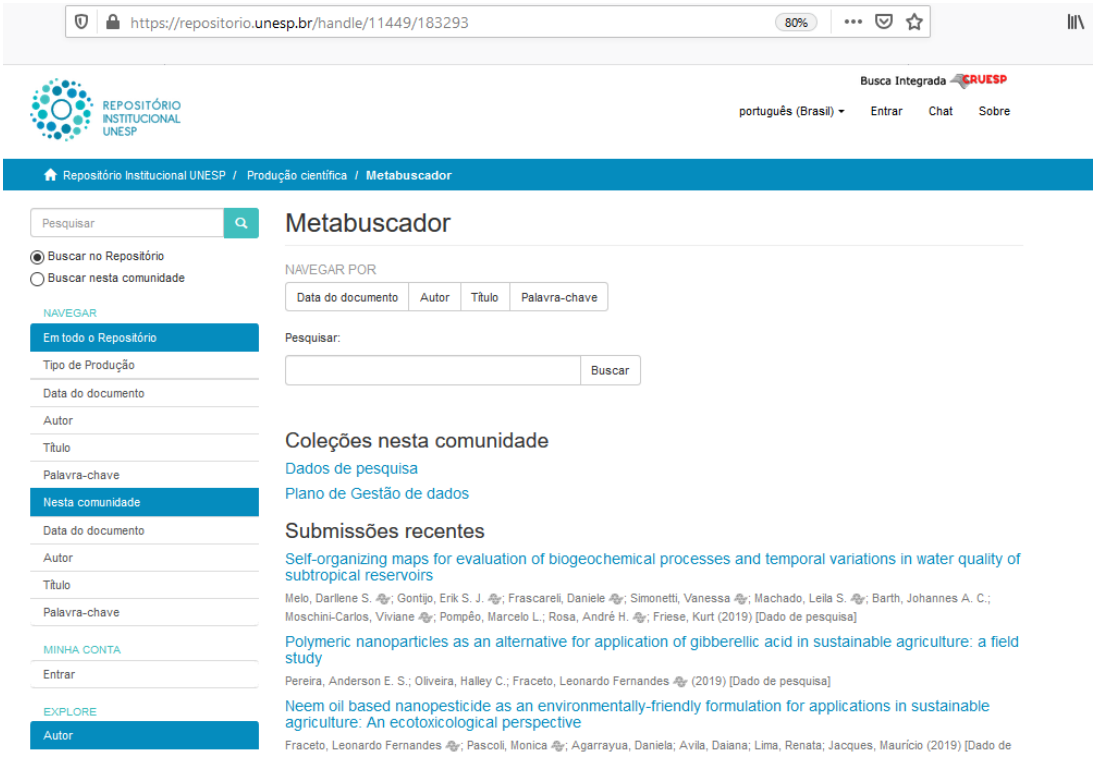
[34]O DMP online é uma ferramenta baseada na Web que oferece suporte a pesquisadores para desenvolver planos de gerenciamento e compartilhamento de dados. A ferramenta é baseada na base de código DMP Roadmap de código aberto gerenciada em parceria com a California Digital Library.
[35] Escritório de Dados Estratégicos Institucionais da UNIFESP.

A instituição criou os Termos Gerais de Uso do Repositório de Dados de Pesquisa e um Guia para os Usuários que orienta os usuários quanto às regras de conduta e *upload* pelo usuário, ou seja, para o autoarquivamento no sistema, além das licenças de uso, *download* do conteúdo, citações e outros.

4.1.1.5 Universidade Estadual Paulista

No ano de 2017, a UNESP firmou convênio de cooperação técnica (011/2018) com a FAPESP, por meio da Coordenadoria Geral de Bibliotecas (CGB). O convênio tem como objetivo compartilhar informações referenciais da produção científica, acadêmica, técnica e artística disponibilizada no Repositório Institucional da UNESP (Figura 12), oriunda de auxílios e bolsas da FAPESP para a coleta e armazenagem na Biblioteca Virtual do Centro de Documentação e Informação (BV-CDI) da própria instituição, bem como a coleta e armazenagem no Repositório Institucional da Unesp dos dados sobre as informações referenciais dos projetos e bolsas indexados na BV/CDi/Fapesp.

Figura 12. Repositório de Dados de Pesquisa da UNESP



Fonte: UNESP (2020)

Em 2018, a UNESP criou e designou seu grupo de trabalho para atuar com a gestão dos dados de pesquisa, Gestão e Repositório de Dados Científicos, pela portaria do Reitor datada de 3 de agosto de 2018 com a finalidade de viabilizar uma infraestrutura computacional que possibilite o armazenamento dos planos de gestão de dados, bem como do repositório de dados científicos gerados pelos pesquisadores da Unesp financiados pela FAPESP. O grupo foi composto por membros da Pró-reitoria de Pesquisa, Coordenadoria de Tecnologia da Informação e Coordenadoria Geral de Bibliotecas (CGB).

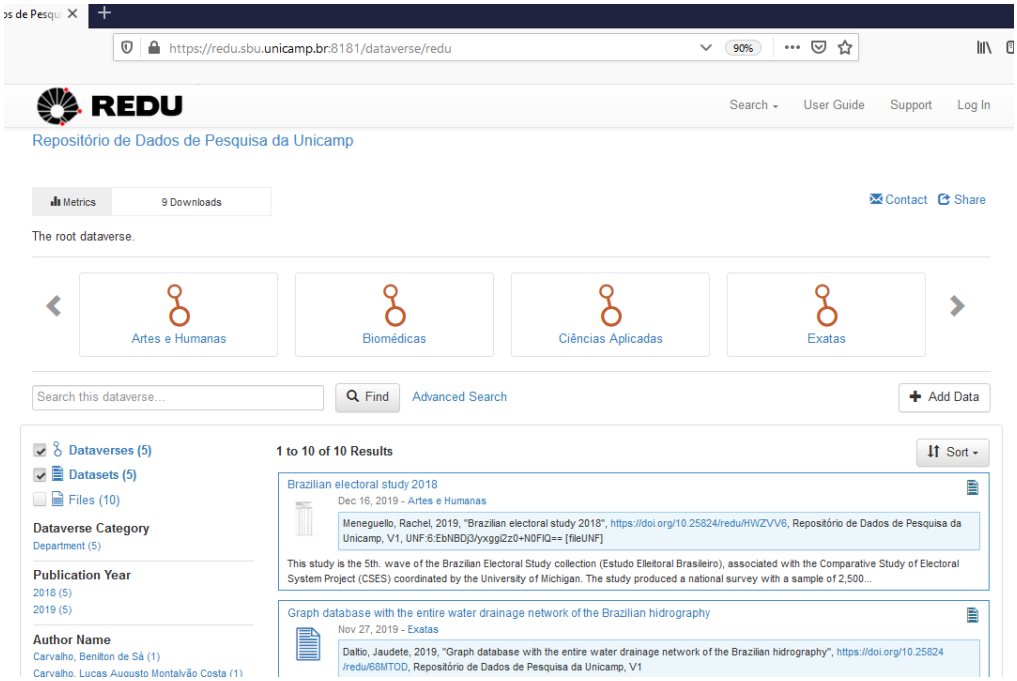
O repositório de dados de pesquisa da UNESP é uma coleção do repositório institucional e está dividido em dados de pesquisa e plano de gestão de dados, conforme apresentado na figura 12. O sistema dispõe apenas de dados de pesquisa e planos de gestão de professores e pesquisadores da instituição financiados pela FAPESP. A inserção destes itens mais os metadados no sistema é feita pelo próprio pesquisador (autoarquivamento), sendo os itens supervisionados. A instituição recomenda o DMP Tool para a criação do plano, dispõem de um Manual de Propriedade Intelectual[36] para orientar os pesquisadores e uma Política de gestão do repositório institucional.

4.1.1.6 Universidade Estadual de Campinas

A universidade foi a pioneira em criar formulários para o gerenciamento de dados científicos e cadastrá-los no diretório DMP Tool. A interface do sistema não contempla muitas informações sobre o processo de criação do Repositório de Dados de Pesquisa da UNICAMP (Figura 13). Identificou-se somente que a instituição recomenda o uso do DMP Tool para a criação dos “planos de gerenciamento de dados, de acordo com as exigências dos principais órgãos de financiamento mundiais” segundo Melis (2018, p. 84). Além disso, exige que os dados científicos que sustentam os resultados nele publicados sejam disponibilizados, de preferência na *Harvard Dataverse.*

[36] PINHEIRO, Patricia Peck. **Manual de Propriedade Intelectual.** São Paulo: UNESP, 2012.

Figura 13- Repositório de Dados de Pesquisa da UNICAMP



Fonte: UNICAMP (2020)

4.1.1.7 Instituto Tecnológico de Aeronáutica

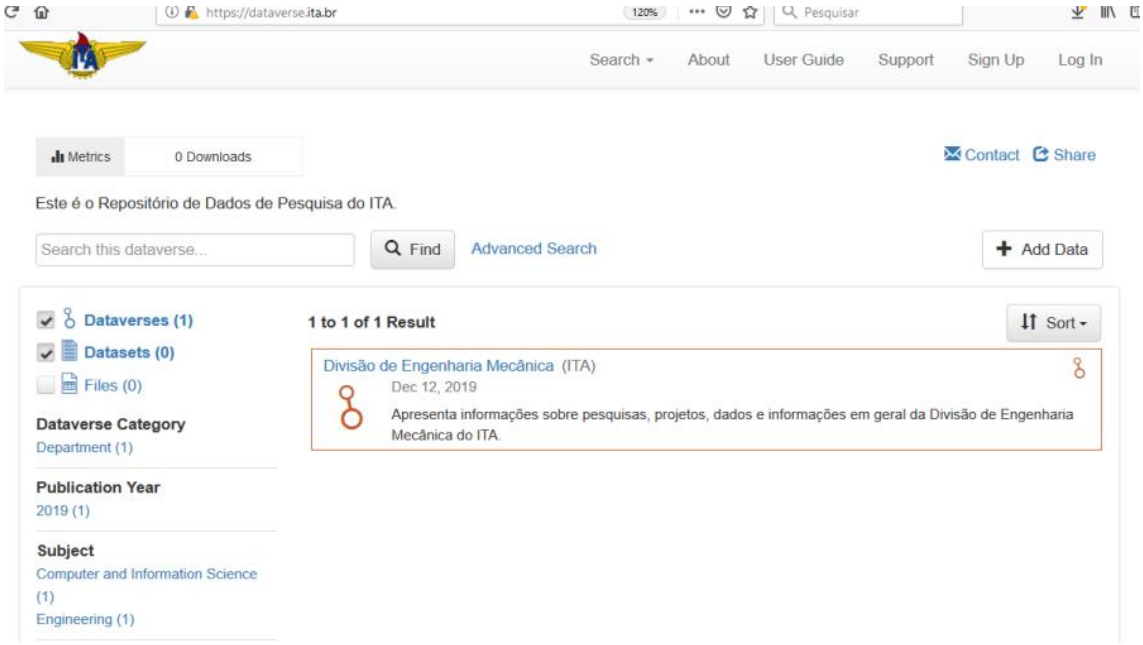
Pela portaria n° 469 de 24 de agosto de 2018, o ITA definiu a composição e as atribuições do seu grupo de trabalho do Plano de Gestão de Dados Científicos. O grupo foi composto por pesquisadores, professores, bibliotecários e técnicos de tecnologia da informação para elaborar uma política de gestão de dados da instituição, analisar e selecionar ferramentas computacionais necessárias.

A instituição realizou diversas atividades de integração com outras instituições para compartilhar experiências com bibliotecários e estudar as melhores práticas para a gestão (alimentação, armazenamento, acesso e preservação dos dados científicos) do repositório.

O Repositório de Dados de Pesquisa do ITA (Figura 14) foi hospedado na Rede de Dados da Biblioteca e utiliza o *software* Dataverse. A instituição fez a instalação da primeira versão para realização da prova de conceito implementando os metadados

já definidos no âmbito de grupo de trabalho de dados da FAPESP, seguido da liberação da versão definitiva.

Figura 14. Repositório de Dados de Pesquisa do ITA



Fonte: ITA (2020)

4.1.1.8 Embrapa Informática Agropecuária

Em maio de 2019 a EMBRAPA instituiu a Política de Governança de Dados, Informação e Conhecimento que estabelece princípios, diretrizes, atribuições e responsabilidades para a gestão de dados, informação e conhecimento, bem como quanto à divulgação de informações relevantes na instituição. Para auxiliar na execução e observância das diretrizes estabelecidas na política, assim como no escopo da segurança da informação, a EMBRAPA criou dois comitês: o Comitê de Governança de Dados, Informação e Conhecimento (CG-DIC) e o Comitê Local de Gestão de Dados, Informação e Conhecimento (CL-DIC) que atua em domínios e contextos específicos. Ambos são compostos por analistas da Secretaria de Desenvolvimento Institucional, representantes do Sistema Embrapa de Bibliotecas, integrantes do Governança de Dados, Informações e Conhecimento, representantes da ouvidoria, da Gerência da Tecnologia da Informação, representantes da Secretaria de Inteligência e Relações Estratégicas (SIRE), representante da Secretaria de Inovação e Negócios (SIN), representante da Secretaria de Pesquisa e

Desenvolvimento (SPD), representantes da Embrapa Informática Agropecuária e representantes de outra Unidade Descentralizada, com experiência nas temáticas contempladas pelos comitês.

O Repositório de Dados de Pesquisa da EMBRAPA (Figura 15) foi criado para suportar e prover diferentes tipos de dados, a exemplo de dados experimentais, geográficos e outros. Em nível de unidade de pesquisa, formou e ampliou uma equipe de Gestão de Dados de Pesquisa (GDP), adotou planos de GDP, implementou um repositório local de dados utilizando o Dspace e Dataverse, exportou algumas coleções para o repositório de dados FAPESP, a exemplo do Digiphatos e realizou a proposição de estrutura tecnológica para o Programa Corporativo de GDP.

Figura 15. Repositório de Dados de Pesquisa da EMBRAPA

Fonte: EMBRAPA (2020)

A instituição também criou o documento “Visão 2014-2034: o futuro do desenvolvimento da agricultura brasileira”, com debates e informações relevantes para os grandes desafios tecnológicos nas diferentes cadeias produtivas

agropecuárias e fornecendo suporte aos planos e às ações estratégicas para a instituição. Este documento apresenta orientações quanto ao acesso, adaptação e desenvolvimento de inovações para a gestão de grandes volumes de dados, incluindo os dados científicos oriundos das pesquisas conduzidas pela instituição. Para reger a questão dos dados científicos disponíveis no repositório de dados da instituição e orientar os usuários quanto ao uso dos mesmos, foi criado um documento denominado de "Termo de Uso Base de Imagens de Sintomas de Doenças de Plantas (PDDB)".

4.1.2 Políticas da Rede

Neste ponto são descritos e discutidos os resultados obtidos mediante análise das políticas da Rede de Repositórios de Dados Científicos do Estado de São Paulo relacionadas ao processo de organização e representação de assunto dos dados científicos dos repositórios analisados.

4.1.2.1 Políticas da Rede da USP

Na USP identificamos a Resolução 7900 de 11 de dezembro de 2019 que estabelece normas para a gestão de dados científicos da instituição, considerando os princípios da Ciência Aberta e como recursos valiosos que devem ser preservados, a partir das exigências de agências de fomento e periódicos para o acesso aberto aos dados e visibilidade da produção científica. A resolução estabelece que a instituição disponibiliza um repositório institucional para armazenar os dados científicos, os quais são elegíveis para o uso pelos docentes da instituição que tenham responsabilidade sobre a publicação desses dados científicos.

As condições para o armazenamento desses recursos são o preenchimento dos metadados necessários de maneira adequada pelo docente e adequação aos aspectos éticos e legais envolvidos na questão. Neste caso o pesquisador faz a solicitação de inserção dos dados científicos e metadados no repositório de dados ao Grupo Gestor de Dados Científicos da instituição, o qual fará a análise segundo

critérios pré-estabelecidos. Após a análise do grupo gestor a biblioteca realiza uma curadoria dos metadados propostos pelo pesquisador, seguidos de inclusão e publicação no repositório de dados da USP. A indexação dos dados científicos é realizada no momento em que o pesquisador envia seu material para o grupo, por meio do preenchimento dos metadados. Assim, o responsável pelos termos atribuídos no metadado assunto é o próprio autor do documento, havendo posteriormente o processo de curadoria realizado pela equipe do repositório.

A resolução também aborda o espaço e o custo de armazenamento, em que o primeiro leva em conta a necessidade do proponente e a disponibilidade de recursos e o segundo poderá ser coberto pela USP ou demandar aporte de recurso por parte do solicitante elegível. Quanto às diretrizes e políticas de uso do repositório por parte dos usuários, preservação digital e curadoria dos metadados, será definida pelos Conselhos de Pesquisa e de Pós-Graduação, portanto em desenvolvimento.

4.1.2.2 Políticas da Rede da UFSCar

Na UFSCar identificamos o Manual de Autodepósito de Dados de Pesquisa que orienta o autodepósito de dados referentes às pesquisas científicas produzidas na instituição, sendo dividido em duas seções que contemplam o depósito de dados de pesquisa com e sem embargo, que consiste no período entre o depósito e a liberação dos dados para acesso público[37]. Os dados disponibilizados no repositório compreendem todos os tipos de recursos informacionais produzidos em qualquer fase do ciclo de vida da pesquisa, viabilizando o acesso, preservação e o uso a longo prazo.

A inserção dos dados de pesquisa e metadados é feita pelo próprio pesquisador por meio do seu número na instituição. Entre os metadados obrigatórios estão o título, autor, URL do Currículo Lattes do autor, unidade da UFSCar, departamento, programa de pós-graduação, data de publicação, descrição, palavras-chave, área do conhecimento[38], idioma e agência de fomento. A representação de assunto no manual

[37] Este período varia de meses a anos e, nestes casos, os dados de pesquisa devem ser depositados, ficando restrito o acesso ao conteúdo até que o embargo esteja vencido.
[38] Conforme a Tabela de áreas do conhecimento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

é contemplada nos metadados descrição e palavras-chave, nos quais o pesquisador tem a obrigatoriedade de informar a finalidade, natureza e escopo do conjunto de dados científicos, assim como as palavras-chaves representativas no mínimo em português e inglês. Outro metadado contemplado, porém não obrigatório, é a descrição do arquivo onde o pesquisador deve fornecer uma breve descrição do mesmo.

Além disso, o repositório institucional em destaque permite relacionar os dados científicos dispostos no sistema com publicações científicas externas ao qual o conjunto de dados está relacionado, em conformidade com o movimento dos dados interligados, assim como atribuir um identificador persistente. Quanto ao *upload* dos arquivos todos os formatos são permitidos. No entanto, a instituição sugere-se as extensões como XML, JSON, CSV, entre outras. Os arquivos depositados (dados científicos e plano de gestão de dados) devem ter tamanho máximo de 5GB, caso ultrapasse deve ser depositado em um repositório externo e referenciado nos metadados. As respectivas licenças dos dados científicos devem ser atribuídas no repositório de dados.

Após o item ser submetido com os seus metadados e plano de gestão de dados pelo pesquisador, a disponibilização *online* do recurso informacional ocorrerá após a sua validação e, assim que aprovado, será disponibilizado automaticamente na interface do RI-UFSCar. Caso o item submetido seja rejeitado o pesquisador receberá um e-mail no qual constará o motivo da rejeição para fins de correção da sua submissão.

4.1.2.3 Políticas da Rede da UFABC

A instituição UFABC está trabalhando na elaboração de sua política de uso do repositório de dados, com previsão para ser finalizado no final do primeiro semestre de 2020. A universidade, como um todo, está se mobilizando para promover a abertura dos seus dados institucionais, sendo instituído o Plano de Dados Abertos 2020-2022 pela portaria n° 451. O documento se configura como orientador para as ações de transparência ativa das bases de dados institucionais de natureza administrativa, não contemplando diretamente os dados científicos.

Sendo o repositório de dados da instituição o canal oficial de divulgação de bases de dados e de estatísticas institucionais, a instituição desenvolveu dois sistemas, um dedicado aos dados científicos e o outro de bases de dados e estatísticas.

4.1.2.4 Políticas da Rede da UNIFESP

Seguindo na mesma linha a UNIFESP desenvolveu o seu Plano de Dados Abertos 2018-2019, que estabelece as suas estratégias para abertura de dados, definindo um conjunto de ações para viabilizar a prática de dados abertos na instituição. A mesma, pela Resolução nº178, dispões sobre a Política de Gestão de Dados Estratégicos Institucionais que estabelece objetivos, princípios, diretrizes e governança dos dados. Porém, a Política de Gestão de Dados de pesquisa da instituição, até a presente pesquisa, não foi concluída, mesmo que está já esteja prevista na abrangente Política de Dados da UNIFESP.

Como instrumento para auxiliar os usuários do repositório de dados, a instituição criou os Termos Gerais de Uso do Repositório de Dados de Pesquisa UNIFESP, disponível no processo de cadastro do usuário no sistema. O termo prevê regras de conduta, *upload* pelo usuário, licenças de uso, citações e um vocabulário de termos utilizados no texto.

Segundo o termo, o usuário registrado ou como convidado não registrado, pode fazer o *download* do conteúdo disponível publicamente. Com acesso controlado ao repositório, o usuário pode obter acesso às funcionalidades de *upload* e *download*. O repositório pode recusar o registro ou cancelar uma conta. Para realizar o *upload* de dados científicos o usuário precisa ser validado pela instituição. O *upload* é feito diretamente pelo usuário, se responsabilizando pelo conteúdo e arquivos submetidos no repositório, assim como a atribuição das licenças de uso e dos metadados.

Fica concedido ao repositório de dados todas as permissões e licenças necessárias para o arquivamento, a preservação e o acesso ao conteúdo depositado, sem restrição ou permissão para disseminar cópias, promover e divulgar o conteúdo, armazenar, traduzir, copiar ou reformatar, incorporar metadados ou documentação. Cabe esclarecer, ainda, que o usuário só pode fazer o *upload* de dados científicos

sem embargo. O depositante não precisa ser o autor da pesquisa, o mesmo pode conceder acesso a outros usuários ao conjunto de dados.

O depositante deve fornecer informações sobre os dados científicos, incluindo, mas não limitado, ao nome do autor, data de publicação, título do conteúdo de dados, descrição do conteúdo e outras informações relacionadas. O usuário pode carregar, enviar, distribuir ou publicar dados, novos conjuntos de dados, informações de metadados, contratos de licença de uso de dados, informações descritivas para páginas de *download* de conjuntos de dados e outras contribuições.

4.1.2.5 Políticas da Rede da EMBRAPA

Na EMBRAPA identificamos a Política de Governança de Dados, Informação e Conhecimento da Embrapa e os Termo de Uso da Base de Imagens de Sintomas de Doenças de Plantas (PDDB). A política dispõe sobre princípios, diretrizes, atribuições e responsabilidades para a gestão dos dados científicos e divulgação de informações relevantes na empresa. A política abrange todas as instâncias organizacionais, secretarias e unidades descentralizadas da instituição. Esta política contempla os dados científicos, dados abertos, dados administrativos e dados pessoais, estes produzidos no exercício das funções da EMBRAPA, sendo propriedade da mesma e gerenciados como ativos corporativos.

A gestão destes dados está alinhada com os objetivos da alta administração da empresa. Os dados científicos são abordados como registros factuais produzidos ou utilizados como fontes primárias para a pesquisa científica e tecnológica e necessários para validação dos resultados. Na perspectiva estratégica, a política prevê implementar, sustentar e monitorar um programa corporativo de gestão de dados científicos e orientar quanto à elaboração de planos de gestão de dados no contexto dos projetos de Pesquisa, Desenvolvimento e Inovação (PD&I). Na perspectiva da interação com o ambiente externo e de negócios, prevê a promoção do uso destes dados para a geração e monitoramento de negócios e estratégias, serviços e produtos digitais.

Quanto ao termo de uso, a política prevê regras de utilização da base de imagens, sendo o usuário toda pessoa física ou jurídica e a EMBRAPA mantenedora

dos direitos de propriedade intelectual e isenta de qualquer responsabilidade sobre eventuais prejuízos. A instituição tem o direito de descontinuar ou alterar a política a qualquer momento sem aviso prévio. O repositório de dados da instituição só permite que os usuários visualizem os dados e façam *download*, na qual estes devem prover os recursos tecnológicos necessários para o acessá-lo. O usuário tem a garantia da confidencialidade de seus dados de navegação no sistema. Na versão atual do repositório de dados da EMBRAPA a inserção dos dados científicos, gestão e a curadoria dos metadados é feita pela própria instituição.

4.1.2.6 Políticas da Rede da UNICAMP, ITA e UNESP

As instituições UNICAMP e ITA não disponibilizam informações e políticas de gestão de dados de pesquisa de seus respectivos repositórios. Por seu turno, a UNESP está atuando no desenvolvimento de guias, termos, materiais de apoio e políticas de acesso, conforme informações presentes em seu repositório de dados. Identificamos que o sistema da UNICAMP é semelhante ao da UNIFESP no sentido de que o usuário, para fazer o *upload* dos dados científicos deve criar uma conta no repositório e ser autenticado pela Equipe do REDU[39] ou entrar com o *login* da Autenticação Central.

4.1.3 Ambiente Digital

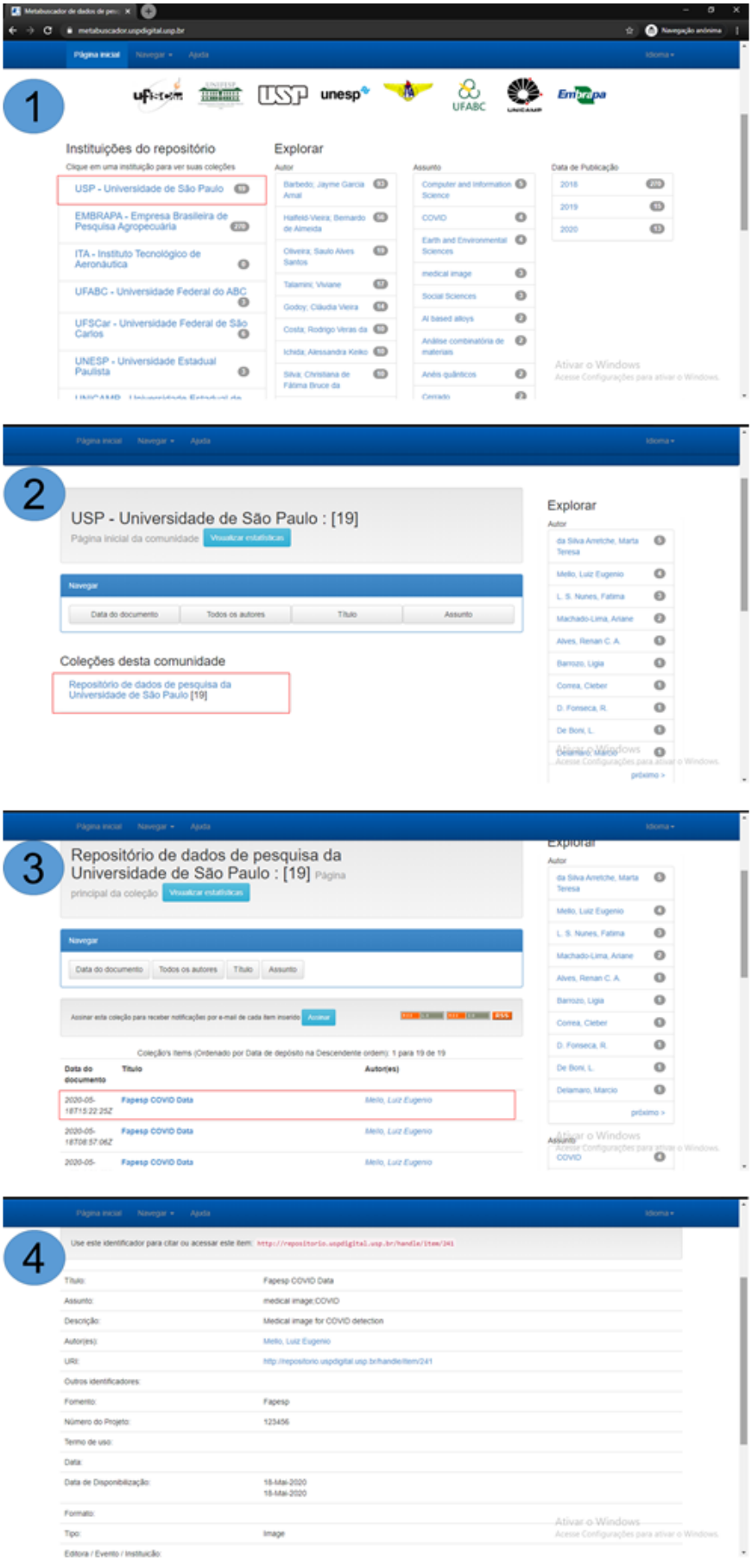
A análise do ambiente digital da Rede de Repositórios de Dados Científicos do Estado de São Paulo dar-se-á inicialmente pelo seu metabuscador, desenvolvido pela USP. Na página inicial do metabuscador são apresentadas as instituições participantes, a finalidade do ambiente, alguns autores, principais assuntos, datas de publicação e uma nuvem de *tags*. As instituições, os autores e os assuntos são apresentados em forma de itens de lista referenciados, ou seja, cada item remete há alguma coisa, direciona há um outro local do servidor, *web* ou item.

[39]Repositório de Dados de pesquisa da Unicamp

Na Figura 16 temos o exemplo destes itens referenciados. Na indicação de número 1 temos a instituição sendo selecionada que encaminha para a indicação 2, esta apresenta a coleção da comunidade. Selecionando o item anterior encaminha para a indicação 3, apresentando os dados científicos ordenado por data de depósito no sistema. Selecionando qualquer um dos dados científicos apresentados, o usuário é encaminhado para a indicação 4 expondo os metadados e os recursos informacionais para *download*.

Esse processo pode continuar a ser executado até chegar ao *layout* da página do repositório de dados da instituição selecionada, cujas especificidades foram descritas em item anterior. Cabe esclarecer, ainda, que o metabuscador foi desenvolvido para reunir e recuperar os dados científicos nos repositórios de dados, sendo que cada instituição criou seus próprios repositórios, interfaces e *layouts*. Este processo pode ser feito em qualquer item do repositório que esteja referenciado, identificado por letras azuis, sublinhados e que recebe clique.

Figura 16. Itens referenciados no metabuscador da FAPESP



Fonte: Metabuscador FAPESP (2020)

No metabuscador os campos existentes para adicionar termos ou palavras-chaves são utilizados para a busca avançada, não existindo um campo que permita que o usuário interaja com o sistema adicionando suas próprias *tags.* Pode se observar que este apenas reúne e recupera os dados científicos que foram adicionados pelos repositórios de dados, portanto não havendo autonomia pelo sistema para modificar e/ou acrescentar algum tipo de informação.

No ambiente de alguns dos repositórios de dados integrantes da Rede de Repositórios de Dados da FAPESP, como UFSCar, UNESP, UNIFESP e UNICAMP, a forma de interação dos usuários com o sistema ocorre quando o mesmo é o autor ou depositante autorizado, e no momento do *upload* quando adiciona os metadados na sua pesquisa, apenas. Este usuário não pode adicionar qualquer informação aos dados científicos de outros pesquisadores. Enquanto que nas demais instituições, o pesquisador envia os dados científicos e os metadados para a instituição, que fará a inserção em seu repositório de dados, não sendo adotado o autoarquivamento nos sistemas. Em todos os repositórios de dados participantes da Rede de Repositórios de Dados da FAPESP não há um campo disponível para os usuários comuns adicionarem *tags*, termos, palavras-chaves ou outras informações.

Na USP, EMBRAPA, UFABC e ITA a indexação dos dados científicos no sistema é feita pelos grupos de trabalhos definidos por cada instituição, com definições prévias sobre os responsáveis em cada um dos grupos pela coleta, curadoria e inserção no repositório de dados[40].

Algo em comum dos ambientes digitais analisados é que estes permitem a navegação por meio de *tags* de autor, assunto, palavras-chaves, data do depósito e da publicação, agência de fomento e outros, com exceção da EMBRAPA que não tem navegação por assunto. No repositório da EMBRAPA não há o campo de assunto e os itens não estão cadastrados como tal, e sim como cultura, título, desordem e pelo autor, conforme descrito na Figura 17.

[40] Na maioria dos repositórios de dados da Rede FAPESP a coleta, curadoria e inserção dos dados é feita pelo Sistema de Bibliotecas das instituições.

Figura 17. Navegação dos dados EMBRAPA

Fonte: EMBRAPA (2020)

4.2 Síntese dos resultados

Nesta seção, apresentam-se os principais resultados obtidos com a avaliação das políticas que regem a organização e o tratamento dos dados científicos no âmbito da Rede de Repositórios de Dados Científicos do Estado de São Paulo, a partir das categorias básicas estabelecidas, as quais foram apresentadas no Quadro 1 desta Dissertação, quais sejam: Responsabilidade; Conteúdo; Aspectos legais; Padrões; Preservação digital; Política de acesso e uso; e Sustentabilidade e Financiamento.

A síntese geral dos resultados visa favorecer um melhor entendimento sobre os pontos evidenciados na pesquisa e aventar as possibilidades da adoção da representação colaborativa de dados científicos na Rede de Repositórios de Dados Científicos do Estado de São Paulo, como uma estratégia para a representação e recuperação por assunto alinhada ao contexto *Big Data*, conforme apresentado no quadro-síntese a seguir (Quadro 9):

Quadro 9. Categorias básicas da Rede de Repositórios FAPESP

CATEGORIA	EMBRAPA	ITA	UFABC	UFSCar	UNESP	UNICAMP	UNIFESP	USP
Responsabilidade	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim
Conteúdo	Dados científicos	Dados científicos	Dados científicos	Dados científicos de docentes, cobrindo todos os tipos de objetos produzidos em qualquer fase do ciclo de vida da pesquisa.	Dados científicos de docentes, cobrindo todos os tipos de objetos produzidos em qualquer fase do ciclo de vida da pesquisa. Os tipos de dados são definidos pelos pesquisadores no plano de gestão de dados.	Dados científicos	Dados científicos	Dados científicos de docentes, cobrindo todos os tipos de objetos produzidos em qualquer fase do ciclo de vida da pesquisa.
Aspectos legais	Não indicado	Não indicado	Definido nos planos de gestão de dados elaborado pelo pesquisador	Definido nos planos de gestão de dados elaborado pelo pesquisador	Definido nos planos de gestão de dados elaborado pelo pesquisador	Não indicado	Definido nos planos de gestão de dados elaborado pelo pesquisador	Definido nos planos de gestão de dados elaborado pelo pesquisador
Padrões	Software Dataverse DOI (Datacite) – sistema de identificação dos dados. Política de Governança de Dados, Informação e Conhecimento da Embrapa.	Recomendação geral: Princípios F.A.I.R. Software Dataverse Handle.Net como sistema de identificação dos dados.	Recomendação geral: Princípios F.A.I.R. Software Dataverse Handle.Net como sistema de identificação dos dados.	Recomendação geral: Princípios F.A.I.R. Software DSpace Padrões de Interoperabilidade sugeridos: XML, JSON e CSV Handle.Net como sistema de identificação dos dados.	Recomendação geral: Princípios F.A.I.R. Software DSpace Handle.Net como sistema de identificação dos dados.	Recomendação geral: Princípios F.A.I.R. Software Dataverse DOI (Datacite) – sistema de identificação dos dados.	Recomendação geral: Princípios F.A.I.R. Software Dataverse Padrão de metadados Dublin Core. Protocolo para coleta de metadados: OAI-PMH. Política de copyright e autoarquivo de editores. Política de Acesso Aberto: em desenvolvimento.	Recomendação geral: Princípios F.A.I.R. Software DSpace com casca própria. Padrões sugeridos: XML, TXT, HTML, PDF, Open Office, CSV, TIFF, PNG, FLAC, WAV, MP3, ZIP Handle.Net como sistema de identificação dos dados.

							Material de apoio aos pesquisadores como manuais, guias, termos e outros: em desenvolvimento. Handle.Net como sistema de identificação dos dados.	
Preservação digital	Recomendações: não violar direitos autorais. Lei de Direitos Autorais nº 9.610, de 19/02/1998 do Governo Federal Brasileiro.	Recomendações: não violar direitos autorais.	Recomendações: não violar direitos autorais.	Recomendações: não violar direitos autorais.	Recomendações: não violar direitos autorais. Política de preservação digital para documentos de arquivo da UNESP. Identificadores digitais: metadados. Manual de Propriedade Intelectual.	Recomendações: não violar direitos autorais.	Recomendações: não violar direitos autorais.	Recomendações: não violar direitos autorais.
Política de acesso e uso	Termo de Uso - Base de Imagens de Sintomas de Doenças de Plantas (PDDB). Acesso livre a toda pessoa física ou jurídica via Internet.	Setor e Política de propriedade intelectual.	Política de uso do RDP (em desenvolvimento).	Todo indivíduo com acesso à Internet pode acessar e usar. Manual de autodepósito de dados de pesquisa. CC BY, CC BY-SA, CC BY-ND, CC BY-NC, CC BY-NC-SA ou CC BY-NC-ND	Dados e planos de acesso livre, respeitando os embargos quando for o caso. Informações de como citar o documento.	Não indicado	UNIFESP – Termos Gerais de Uso do Repositório de Dados de Pesquisa. Guia para os Usuários do RPD. Licença padrão CC BY.	Resolução 7900/2019. CC BY, CC BY-SA, CC BY-ND, CC BY-NC, CC BY-NC-SA ou CC BY-NC-ND
Sustentabilidade e Financiamento	Recursos do Governo Federal, iniciativas privadas, organizações de pesquisa e inovação e sociedade civil.	Não indicado	Financiamento do programa de pesquisa Horizonte 2020 da União Europeia	Não indicado	Recursos da própria instituição	Não indicado	Não indicado	Não indicado

Fonte: Dados da pesquisa

5 CONSIDERAÇÕES FINAIS

Diversas áreas do conhecimento científico têm estudado *Big Data*, suas implicações, desafios, benefícios, *modus operandi* e outros aspectos intrínsecos a essa nova configuração do ambiente digital. Neste processo de envolvimento com o esse fenômeno contemporâneo, muitas dessas áreas têm desconstruído seus paradigmas, dogmas, cânones e se reinventado enquanto ciência.

A Ciência da Informação tem enfrentado este desafio, tendo como resultado desse movimento diversas publicações nacionais e internacionais dedicadas à questão, dentre as quais encontra-se a representação de grandes volumes de dados e informações em um ambiente dinâmico, heterogêneo e descentralizado como a *Web*, ao passo que os sistemas tradicionais de organização do conhecimento mostram-se cada vez mais insuficientes (HJORLAND, 2012; 2013). Nesta perspectiva, estudiosos do tema na Ciência da Informação ressaltam que o *Big Data* tem impactado, direta ou indiretamente, as técnicas tradicionais de análise de assunto, análise documentária ou análise temática, impossibilitando o emprego da indexação de assunto da forma tradicional, mediante tratamento individual e intelectual dos registros.

Desta forma, uma das estratégias para minimizar este problema se converte na inserção dos usuários no processo de representação de assunto, atuando como colaboradores nos sistemas de informação ao lado dos bibliotecários e cientistas da informação. As indicações teóricas apresentadas nesta pesquisa apontam a Folksonomia como uma das possibilidades de se conduzir a organização e o tratamento de recursos digitais na era do *Big Data*, visto que as questões a serem discutidas no âmbito do fenômeno não se concentram mais nos aspectos tecnológicos, mas na forma de prover a organização no ambiente digital.

Em alguns estudos a adoção desta estratégia gerou um número maior de dados relacionados ao recurso informacional, maior variedade de termos, diversos contextos representados e resultados das buscas mais relevantes (VIANA; DAL'EVEDOVE; GRACIOSO, 2019). A Folksonomia se estabelece como uma estratégia eficiente para a representação e recuperação por assunto em ambientes digitais como observado por Viana, Arakaki e Dal'Evedove (2019) ao discutir "A contribuição da folksonomia para a construção de metadados". No referido artigo, as

autoras constatam que os metadados quando associados às contribuições da folksonomia enquanto recurso categorial coletivo é possível impulsionar os resultados positivos de seu uso em ambientes digitais. Desta forma:

> a representação colaborativa em plataformas digitais possui grande importância no processo de construção semântica e classificatória dos recursos informacionais à medida que incorpora representações subjetivas em estruturas de formas associadas a modelos de representação formal. (VIANA; ARAKAKI; DAL'EVEDOVE, 2019, p. 521).

Os repositórios digitais têm se consolidado como bases de dados de alto valor científico, confiáveis, estáveis e acessíveis, tanto para a comunidade científica ligada diretamente ao sistema, quanto para a população em geral. Alguns destes repositórios podem disponibilizar recursos informacionais que exigem o uso de metadados que descrevem o contexto de produção e as ferramentas utilizadas para criar, armazenar, adaptar e analisar o recurso informacional que está sendo disponível, a exemplo dos repositórios de dados científicos. Os repositórios de dados científicos, além de oferecerem uma base tecnológica para a contextualização desses tipos de recursos informacionais, tornam-se cada vez mais essenciais para a representação de assunto, recuperação, colaboração dos pesquisadores quanto aos metadados, compartilhamentos e outros. Os dados científicos disponíveis nestes ambientes digitais necessitam de um ecossistema que possibilite o gerenciamento adequado, assim como a contribuição dos pesquisadores na descrição dos metadados.

Nesta perspectiva, objetivou-se investigar as práticas de organização e tratamento da informação em repositórios de dados científicos, no intuito de contribuir com subsídios para a representação colaborativa de registros científicos provenientes de atividades de investigação. Mediante pesquisa exploratória na literatura especializada de Ciência da Informação, foram conceituadas as potencialidades da representação colaborativa da informação em tempos de *Big Data*.

A pesquisa revelou estudos que sinalizam para a atualização ou adequação dos processos e instrumentos tradicionais de organização do conhecimento e representação da informação, visto que os mesmos não incorporam as transformações trazidas com o avanço das tecnologias digitais. No contexto de *Big Data*, há uma mudança na forma de organizar e na forma de acessar os recursos

informacionais, sendo a Folksonomia uma possibilidade mais viável, econômica e prática para representação e recuperação da informação no ambiente digital.

Constatou-se, ainda, que a Ciência da Informação enfrenta desafios em incorporar sistemas de classificação construídos por especialistas e solicitar contribuições de amadores via plataforma *Web* 2.0, tendo em vista que esta ação rompe com uma tradição secular da área. Todo processo de mudança sofre resistência e o campo de conhecimento da Ciência da Informação está em processo de compreensão dos fenômenos emergentes. Movimento necessário para mudar os conceitos e os procedimentos de um campo do saber, formando uma nova base para a prática da ciência (KUHN, 1970).

O impacto do *Big Data* tem realçado as discussões sobre os processos de análise e representação de grandes volumes de dados no ambiente digital. Discutir o *Big Data* no contexto da Ciência da Informação e, portanto, na Organização do Conhecimento, é observar um panorama de confluência e troca de teorias e práticas entre as áreas. O *Big Data* necessita de uma ciência que forneça parâmetros para a análise dos dados e produza conhecimento a partir deles. Apesar de a Ciência da Informação estudar os processos relacionados à informação, a mesma corre o risco de não ter autonomia para enfrentar as implicações geradas pelo fenômeno.

Verificou-se que os estudos nacionais que relacionam a representação colaborativa da informação ao *Big Data* são incipientes, resultado que pode ser reflexo da recente manifestação dessa prática nas atividades sociais e, por decorrência, no escopo investigativo da área e também pela incipiente presença de estudos sobre folksonomias nas pesquisas acadêmicas. Este último ponto foi observado na pesquisa de Viana, Dal'Evedove e Tartarotti (2019), na qual investigaram a inserção e a frequência com que a Folksonomia foi objeto de estudo na pós-graduação em Ciência da Informação no Brasil, na modalidade *stricto sensu*. Identificando pouca produção e discussão no âmbito de disciplinas, teses e dissertações sendo que dos 17 programas de pós, apenas 6 abordavam questões relacionadas as folksonomias. Desta forma, considera-se que uma maior discussão do tema nos programas poderia trazer contribuições efetivas para o avanço do temário na área.

No geral, o envolvimento com a literatura especializada de Folksonomia revela uma preocupação dos pesquisadores em conhecer e analisar as estratégias de indexação social utilizadas em diferentes redes sociais ao longo dos anos. As

pesquisas mais atuais discorrem sobre a hibridização ou concomitância da Folksonomia com vocabulários controlados, tesauros e taxonomias para fins de representação, organização e recuperação da informação. Apesar das pesquisas brasileiras trazerem importantes discussões e recomendações relativas à Folksonomia, o *Big Data* não é mencionado ou considerado. De outro modo, pesquisadores no contexto internacional da Ciência da Informação têm discutido a interseção da Folksonomia com o *Big Data*, mas ainda de forma tímida. Nota-se que a área pouco tem estudado o uso de ferramentas colaborativas no contexto do *Big Data*, resultado que revela a necessidade de um maior envolvimento dos cientistas da informação com o tema.

A análise da Rede de Repositórios de Dados Científicos do Estado de São Paulo revela que o sistema se apresenta como uma importante e pioneira iniciativa para o compartilhamento e preservação de dados científicos oriundos das pesquisas conduzidas no Brasil. Observou-se que o sistema em destaque permite conjugar dados e as informações das oitos instituições participantes, além das características, especificações e políticas de cada uma. A estratégia da FAPESP em construir a rede de repositórios de dados científicos pode ser comparada há uma estrutura construída em Lego, na qual cada instituição tem sua responsabilidade e contribui com uma parte para a construção e manutenção do sistema.

Existe na rede repositórios digitais e/ou institucionais que tiveram seus serviços expandidos para abranger os dados científicos e, ainda, repositórios voltados especificamente para o depósito de dados científicos. A análise da rede a partir dos eixos categorias básicas, políticas da rede e ambiente digital, viabilizou uma visão holística do sistema da FAPESP. Compreendendo que o insumo básico para o desenvolvimento de qualquer rede, repositórios e outros sistemas, em que os mesmos permitam dialogar com as contribuições do usuário, estará contemplada nestas categorias básicas e políticas da rede. O ambiente digital será a consolidação ou representação de como aquele sistema está estruturado.

Em algumas políticas da rede existem elementos dedicados ao ciclo de vida dos dados científicos, contemplando as fases de coleta, armazenamento, recuperação e uso, em consonância com as abordagens teóricas da Ciência da Informação. A maior parte das instituições esclarecem em suas políticas o que são os dados científicos e quem pode atuar como pesquisador responsável. Ainda, constatou-se

que os repositórios de dados vinculados à rede estão em processo de adequação aos princípios FAIR, utilizam *softwares* e padrões que permitem a interoperabilidade, possuem sistemas de identificação dos dados e protocolo para a coleta de metadados. A UNESP, UFSCar e a UNIFESP são as únicas instituições da rede que apresentam políticas e permitem o autoarquivamento dos dados científicos pelos pesquisadores.

É comum a todos os repositórios da rede que a responsabilidade pela curadoria dos metadados, inserção tanto dos dados quanto dos metadados no sistema, elaboração de políticas, manuais e a gestão dos repositórios é do sistema de bibliotecas. Nos repositórios analisados há duas formas de representação de assunto dos dados científicos. Uma se consolida na inserção dos itens (arquivo do dado de pesquisa, metadados, descrição, autores, palavras-chave, licenças e etc) pelo próprio pesquisador no sistema do repositório, ou seja, a relação com o sistema é direta, sem intermediários. Na outra, o pesquisador envia esses itens para a equipe responsável pelo repositório e estes inserem no sistema, sendo a equipe responsável o elo entre o pesquisador e o repositório de dados.

Na primeira forma de representação de assunto, alguns repositórios de dados integrantes da rede, por meio de suas políticas, solicitam as contribuições de seus pesquisadores para armazenar e representar, por meio de metadados e termos descritivos, os assuntos dos dados científicos. A contribuição permitida é apenas do pesquisador responsável, porém já indica o valor desta contribuição para a instituição responsável pelo repositório, mesmo que está ainda não tenha se consolidado em práticas de representação colaborativa da informação. Visando a melhoria do sistema e adequação às práticas da *Web* social, a instituição poderia permitir a contribuição não só do pesquisador responsável como também de outros pesquisadores da mesma linha, grupo ou área de estudo e investigação. Desta forma, a instituição conseguiria implementar algumas ações de Folksonomia entre os seus próprios pesquisadores.

Na segunda forma de representação de assunto, observa-se a confluência de duas percepções: a do pesquisador com os metadados da pesquisa e a do profissional responsável pela inserção, representação e curadoria dos dados no sistema. Nesta a representação de assunto não se configura exclusivamente como unidirecional, mas parte do princípio de colaboração entre os envolvidos, contribuindo com a construção

de bons metadados e adotando abordagens *top-down* a fim de atingirem o equilíbrio certo entre duas abordagens aparentemente opostas.

Enquanto algumas instituições voltam-se para os seus repositórios e planos de gestão de dados, outras reconheceram a importância dos dados científicos, elevando-os nas discussões e políticas para além da organização, como também um alinhamento estratégico da instituição. Mesmo com tais atitudes, a Folksonomia não foi considerada para a organização e recuperação dos dados científicos da rede, ausentes nas categorias básicas, nas políticas e nos ambientes digitais. Como podemos observar nos resultados desta pesquisa, o sistema da rede tem potencial para agregar a representação colaborativa da informação por possuir uma infraestrutura tecnológica, padrões de interoperabilidade, protocolos de coleta de metadados, políticas de autoarquivamento e sua gestão ser conduzida por profissionais ligados ao campo científico e profissional da Ciência da Informação.

Observa-se que a rede de repositórios de dados da FAPESP tem um grande potencial e ambiente propício para incorporar a representação colaborativa em seu sistema, pois a mesma já dispõem de infraestrutura. Porém a rede não utiliza deste sistema flexível para indexar, classificar e recuperar os dados científicos de seus repositórios, com exceção da UFSCar, UNESP e UNIFESP que permite o *upload* e a indexação dos dados pelo usuário-autor.

Acredita-se que o compartilhamento de informações é essencial para agilizar e maximizar os avanços da ciência, evitando trabalho redundante, facilitando a reprodutibilidade dos experimentos e, ainda, contribuindo para que pessoas com motivações e interesses semelhantes, mas separadas pela distância geográfica, reúnam-se em torno de projetos comuns (WILBANKS; BOYLE; REYNOLDS, 2006). Neste sentido, considera-se que os repositórios de dados científicos são sistemas ideais para a prática folksonômica, tanto em suas políticas como ambiente digital, mesmo que ainda limitados, pois já dispõe de infraestrutura que possibilita a contribuição dos pesquisadores e há um processo de curadoria dos metadados inseridos pelos usuários por especialistas. Com efeito, tem-se ampliada as possibilidades de os repositórios de dados científicos atuarem como sistemas mais democráticos, inclusivos e representativos.

A ausência de discussões teóricas sobre a representação colaborativa no contexto de repositórios de dados científicos como alternativa para a organização e

tratamento dos registros científicos provenientes de atividades de investigação pode ser uma das razões pela falta de adesão pelos sistemas pertencentes a Rede de Repositórios de Dados Científicos do Estado de São Paulo. Como alternativa para a questão, tem-se a necessidade de discussões complementares que cerquem os temas contemplados na presente pesquisa tendo como referência outras redes ou repositórios de dados científicos nacionais e internacionais para o avanço deste eixo investigativo na Ciência da Informação. Estas pesquisas podem identificar repositórios de dados científicos que adotam a representação colaborativa como prática para a organização e tratamento de dados científicos e oferecerem subsídios complementares aos aqui apresentados, assim como revelarem novas perspectivas de discussão para o tema. Da mesma forma, há a necessidade de investigar junto aos gestores das instituições e os gestores dos Sistemas de Bibliotecas pesquisados a sistematização, aplicação e discussão das temáticas abordadas nesta pesquisa.

Considera-se um dos impactos positivos do uso da Folksonomia nos repositórios de dados científicos a possibilidade de reutilização dos metadados, dos termos descritivos, das palavras-chave e dos nomes dos autores para a indexação e representação de outros dados. Visto que estes repositórios precisam indexar, classificar e recuperar extensas coleções de dados e inferir resultado de busca satisfatório para o usuário.

Depreende-se algumas considerações a respeito da inserção das Folksonomia nos sistemas acadêmicos, como repositórios, sistemas de gestão de periódicos, sistemas de gestão de bibliotecas, não serem naturais como o uso e a inserção deste nas redes sociais. Do ponto de vista da autora desta pesquisa, o *layout* do ambiente digital é um dos fatores. É comum observar que nestes sistemas acadêmicos não há uma preocupação em apresentar um *design* atrativo, uma estética agradável, uma facilidade e eficiência, praticidade e uma boa experiência de navegação, nem aos usuários e nem aos profissionais que gerenciam o sistema. Outra situação é que os sistemas são construídos para os usuários consumirem as informações e não para interagirem, os sistemas são fechados, a maioria destes ambientes também não são responsivos, ou seja, o ambiente digital não é projetado para se adaptar a qualquer tipo de resolução (*desktops*, *tablets*, dispositivos *mobile* e outros), sem distorções.

Observa-se o tipo de usuário dos sistemas acadêmicos, estes são mais objetivos e talvez até preciso em suas buscas. Buscam informações relevantes e

fidedignas, sendo o objeto de pesquisa a informação científica, e na maioria os usuários são do ambiente acadêmico, ou seja, são professores, alunos de graduação ou pós-graduação e pesquisadores no geral.

Este último aspecto pode ser considerado um fator relevante e positivo para a inserção das Folksonomia nos sistemas acadêmicos, mais precisamente nos repositórios de dados, pelo observado no perfil do público. Desta forma a construção de *tags* por estes usuários se tornaria mais representativa, mais exaustiva, mais consistente e com maior qualidade, diminuindo a questão de ausência de significados e ambiguidade das tags. Deste modo Lancaster (2007, p. 7) afirma que “à medida que se aumenta a extensão da representação também se aumenta a recuperabilidade do documento”, possibilitando muito mais pontos de acesso. Apesar de não haver uma transição natural da Folksonomia para os sistemas acadêmicos como nas redes sociais, é possível inseri-las nos sistemas acadêmicos considerando o seu público, atenuando a indexação dos recursos pelos usuários, melhorando a experiência do mesmo e possibilitando meios para apoiar o seu uso personalizado.

Da mesma forma que o aspecto apresentado pode ser positivo também pode ser negativo, pelo fato dos mesmos não estarem acostumados com esse tipo de interação nos sistemas acadêmicos devido a rigidez e a oclusão destas estruturas, podendo levar a inexistência de análise e assimilação de termos descritivos, ou seja, a não atribuição de *tags*. Portanto, como alternativa à questão, sistemas que adotarem a representação colaborativa devem promover treinamentos, de modo a capacitar profissionais e usuários. Capacitar no sentido de instruir quanto ao conceito de Folksonomia, a importância, os pontos positivos e negativos, como inserir palavras-chave relevantes, o impacto de boas palavras-chave, e principalmente inserir o usuário neste processo decisório para gerar o senso de pertencimento e promover engajamento. Desta forma o usuário vai tanto se sentir parte da construção do sistema acadêmico como vai possibilitar o crescimento do mesmo, crescimento que pode refletir no número de acesso, no tempo que o usuário passa acessando o sistema, no número de recursos indexados e de tags adicionadas, na redefinição dos tipos de usuários do sistema e outros.

A rede de repositórios de dados científicos da FAPESP reúne 8 instituições distribuídas pelo estado de São Paulo, destas 5 universidades estão entre as 20

melhores do Brasil[41] (Ranking Universitário da Folha, 2019) e as mesmas estão entre as melhores do mundo[42]. As instituições acabam agregando pessoas de várias regiões do Brasil e do mundo em suas disciplinas, cursos de graduação e pós-graduação, professores e técnicos administrativos. Pessoas com diferentes origens culturais, sociais, econômicas, políticas e intelectuais.

Sob estas circunstâncias, os esquemas de indexação e classificação tradicionais se tornam caros, perdem a capacidade de corresponder ao modo de pensar e organizar o mundo pelo ponto de vista dos usuários e não consideram as diversidades. Seria incoerente pensar em representar todas estas diferenças e diversidades em estruturas e padrões rígidos e inflexíveis de organização e compartilhamento da informação. Assim sendo, as folksonomias fornecem uma abordagem democrática e colaborativa de indexação, acesso e recuperação da informação concretizada pelos próprios indivíduos. Neste processo os diferentes aspectos sociais, culturais, econômicos, políticos e intelectuais dos usuários podem ser contemplados por meio das *tags.* Acreditamos que, enquanto profissionais que zelam pelo acesso, compartilhamento, qualidade e fidedignidade da informação, devemos explorar maneiras em que as muitas interpretações possam ser incorporadas em um sistema de indexação.

No transcurso deste estudo, mais questionamentos foram suscitados do que respondidos. Dentre eles, embora a Folksonomia se apresente como a salvação do conflito de que o usuário agora pode participar do processo de indexação e organização da informação junto com especialistas em ambientes digitais, será que ela é de fato oportuna para todos os sistemas, para todas as propostas de compartilhamento da informação? Porque se o sistema não abre a sua estrutura para permitir a adição de *tags* pelos usuários, ele tem uma lógica também de construção, de organização, de indexação e de compartilhamento.

No que diz respeito aos padrões de metadados, em que medida os padrões preveem metadados com campos para indexação colaborativa? Porque eventualmente cada repositório acaba se sustentando nas estruturas de metadados, portanto esta infraestrutura de metadados permite cruzar as perspectivas do usuário

[41] As universidades são USP, UNICAMP, UNESP, UFSCar e UNIFESP. Disponível em: https://ruf.folha.uol.com.br/2019/ranking-de-universidades/principal/. Acesso em: 07 de maio. 2020

[42] Ranking das 1000 melhores universidades do mundo. Disponível em: http://cf.datawrapper.de/eKpYn/1/. Acesso em: 07 de maio. 2020

e do especialista? Visto que a opção saudável e completa seria a adoção de metadados que permitam esse tipo de entrada do usuário.

Portanto para algumas naturezas informacionais ou tipologias documentais, ou até para um determinado público talvez não seja o caso de usar este sistema folksonômico. Desta forma, podemos considerar que a inserção da representação colaborativa na rede de repositórios da FAPESP neste momento, primeiro semestre de 2020, não seja viável, talvez seja o caso de concentrar o movimento de indexação entre os pesquisadores da comunidade acadêmica.

Parte dos repositórios está desenvolvendo suas políticas, suas estruturas de *hardware* e *software*, seus ambientes digitais e a rede ainda está sendo aprimorada. Outro fator observado é que em algumas instituições os dados científicos são considerados ativos da mesma, enquanto que em outras as discussões a respeito parecem limitadas ou levadas em consideração devido a certa exigência. Porém é interessante que a representação colaborativa da informação comece a ser discutida nas instituições, tendo em vista que este estudo piloto, das possibilidades de implantação da representação colaborativa na rede, responde a indicações teóricas da Ciência da Informação e a crescente criação de sistemas neste formato "aberto" na *Web*. Também é importante sinalizar sobre a importância da elaboração de políticas de indexação, de modo a contemplar as limitações ou viabilidades da adoção da representação colaborativa nesses sistemas e, por conseguinte, na rede. De igual forma, que eventos fossem promovidos para discutir este sistema colaborativo, que os usuários fossem requeridos à prestar suas opiniões, que um projeto piloto fosse implantado para testar, levantar e confrontar hipóteses e, ao término, comprovando a eficácia do sistema que fosse contemplada nas políticas e na infraestrutura tecnológica dos repositórios participantes da rede, assim como da própria FAPESP.

Finalizamos esta pesquisa considerando que a Folksonomia é um sistema de organização do conhecimento perante ao contexto do *Big Data*. As folksonomias surgiram para sanar uma necessidade dos usuários das redes sociais em organizar suas informações nestas redes, como Flickr, Delicious e outros. Com o passar dos anos elas passaram a ser uma ferramenta necessária, expandindo o seu uso para além das redes sociais. Esta expansão possibilitou-se, principalmente, pelo o facilitado e rápido manuseio, criação e compartilhamento de *tags*, pela ligação que as

tags criadas pelos usuários fazem com os recursos (imagem, texto, vídeo, etc) e com outros usuários (formando o tripé da folksonomia - *tag*, usuário e recurso).

A Folksonomia é um sistema de organização do conhecimento que permite tanto a indexação feita por especialistas como aquela feita por usuários amadores. E quando esse processo permite reunir esses diferentes personagens na indexação, os termos representativos dos diferentes recursos informacionais se tornam mais exaustivos, mais abrangentes, podem se tornar mais representativos, entre outros. E como todo os sistemas, a folksonomia tem um processo de atualização, correção e curadoria das *tags*, na qual o próprio usuário realiza quando propõe novas *tags*, corrigi *tag* já atribuída, quando os usuários deixam de usar determinadas *tags* e outras situações. Portanto, a folksonomia se torna um sistema de organização prático, fácil (porém não simples), rápido, acessível e de bom engajamento dos seus usuários (participação). E a sua utilização vai depender dos propósitos, dos usuários e da visibilidade que as instituições desejam alcançar.

REFERÊNCIAS

ALBAGLI, Sarita; CLINIO, Anne; RAYCHTOCK, Sabryna. Ciência Aberta: correntes interpretativas e tipos de ação. **Liinc em Revista**, Rio de Janeiro, v. 10, n. 2, p. 434 - 450, nov. 2014. Disponível em: http://revista.ibict.br/liinc/article/view/3593/3072. Acesso em: 06 jan. 2020.

AQUINO, M. C. A Potencialização da Memória Coletiva através do Hipertexto na Web 2.0. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 30., 2007, Santos. **Anais...** Santos: Sociedade Brasileira de Estudos interdisciplinares da Comunicação, 2007. Disponível em: http://www.intercom.org.br/papers/nacionais/2007/lista_area_NP-TI.htm. Acesso em: 03 ago. 2019.

AUCKLAND, M. Re-skilling for research: an investigation into the role and skills of subject and liaison librarians required to effectively support the evolving information needs of researchers. London: Research Libraries UK, 2012. Disponível em: http://www.rluk.ac.uk/content/re-skilling-research. Acesso em: 25 jan. 2020.

AVENTURIER, Pascal. Princípios FAIR: critérios de qualidade para dados de pesquisa. Disponível em: https://publicient.hypotheses.org/1456. Acesso em: 04 dez. 2019.

AVENTURIER, Pascal; ALENCAR, Maria de Cleófas Faggion. Os desafios dos dados de pesquisa abertos. **RECIIS**, Rio de Janeiro, v. 10, n. 3, p. 1-19, jul./set.2016. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1069/pdf1069. Acesso em: 04 dez. 2019.

BAPTISTA, A. A.; CATARINO, M. E. Folksonomia: um novo conceito para a organização dos recursos digitais na Web. **DataGramaZero** - Revista de Ciência da Informação - v. 8, n. 3, 2007.

BARACHO, R. M. A. et al. O caminhar da Ciência da Informação e o XV Encontro Nacional de Pesquisa em Ciência da Informação. **Perspectivas em Gestão & Conhecimento**. v. 4, out. 2014. p. 198-211.

BARDIN, Laurence. **Análise de conteúdo**. 4. ed. Lisboa: Edições 70, 2011.

BARRETO, A. A. A aventura de perceber significados. **Datagramazero**-Revista de informação. v. 15, n. 3, jun. 2014.

BARROS, Léa Maria de Souza. **A folksonomia como prática de classificação colaborativa para a recuperação da informação**. Dissertação (mestrado), Universidade Federal do Rio de Janeiro, 2011. Programa de Pós-Graduação em Ciência da Informação/IBICT. Rio de Janeiro

BASE DE DADOS REFERENCIAL DE ARTIGOS DE PERIÓDICOS EM CIÊNCIA DA INFORMAÇÃO. Disponível em: http://www.brapci.inf.br/. Acesso em: 09 set. 2018.

BORKO, H. Information Science: What is it? American Documentation, v.19, n.1, p.3-5, Jan. 1968.

BOYD, D; CRAWFORD, K. Critical questions for Big Data. **Information, Communication & Society,** v. 15, n. 5, p. 662-679. 2012. Doi: 10.1080/1369118X.2012.678878

BRADLEY, Jean-Claude. **Interview with Jean-Claude Bradley. The Impact of Open Notebook Science by Richard Poynder.** September, 2010. Disponível em: http://www.infotoday.com/it/sep10/Poynder.shtml#top.Acesso em: Acesso em: 06 jan. 2020.

BRADLEY, Jean-Claude. **Shining a light on chemical properties with Open Notebook Science and open strategies.** Lecture at ACS Symposium on August 20, 2012. Disponível em: https://www.youtube.com/watch?v=7mK6SY-jXqk. Acesso em: Acesso em: 06 jan. 2020.

BRIGIDI, Fabiana Hennies. **Indexação híbrida**: vocabulário controlado e folksonomia. 199 p. Dissertação (Mestrado) - Universidade do Estado de Santa Catarina, Florianópolis, SC, 2016.

BROOKS, T. A. Where is meaning when from is gone? Knowledge representation on the web. **Information Research**. v. 6, n. 2, jan. 2001.

CAPURRO, R. Epistemologia e ciência da informação. In: **V Encontro Nacional de Pesquisa em Ciência da Informação**. Belo Horizonte, Brasil, 10 de Novembro de 2003. Tradução de Ana Maria Rezende Cabral, Eduardo Wense Dias, Isis Paim, Ligia Maria Moreira Dumont, Marta Pinheiro Aun e Mônica Erichsen Nassif Borges. 42 p.

CAPURRO,R.; HJORLAND,B. **The concept of Information**. 2003. Disponível em: www.capurro.de/infoconcept.html. Acesso em: 03 Mar. 2018.

CATARINO, M. E.; BAPTISTA, A. A. Folksonomia: um novo conceito para organização dos recursos digitais na Web. **Data Grama Zero**, Brasília, v. 8, n. 3, jun. 2007. Disponível em: http://www.dgz.org.br/jun07/F_I_aut.htm. Acesso em: 09 set. 2018

CATARINO, M. E.; BAPTISTA, A. A. Folksonomias: características das etiquetas na descrição de recursos da web. **Informação & Informação**, Londrina, v. 14, n. esp., p. 46-67, 2009. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/3234 Acesso em: 09 set. 2018

CINTRA, A. M. M. Determinação do tema de pesquisa. **Revista Ciência da Informação, Ci.Inf.**, Brasília, v. 11, n. 3, p. 13-16, 1982.

CONEGLIAN, C. S.; GONÇALVEZ, P. R. V. A.; SANTAREM SEGUNDO, J. E. O profissional da informação na era big data. **Encontros Bibli**: revista eletrônica de Biblioteconomia e Ciência da Informação. v. 22, n. 50, set./dez. 2017. p. 128-143.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E.; SANTANA, R. C. G. Big data: fatores potencialmente discriminatórios em análise de dados. **Em Questão.** 2016. p. 25.

CORDEIRO, D. et al. Da ciência à e-ciência: paradigmas da descoberta do conhecimento. **Revista USP**, n. 97, 2013. p. 71-81. Disponível em: https://doi.org/10.11606/issn.2316-9036.v0i97p71-81. Acesso em: 09 set. 2018

CURTY, R. G. **Beyond data thrifting**: an investigation of factors influencing research data reuse in the social sciences. 2015. 295f. Tese (Doctor of Philosophy in Information Science 201 and Technology) - School of Information Studies, Syracuse University, Syracuse, 2015. Disponível em: http://www.ijdc.net/index.php/ijdc/article/view/401. Acesso em: 28 set. 2019.

CURTY, R. G.; AVENTURIER, P.. O paradigma da publicação de dados e suas diferentes abordagens. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais …** Marília: Unesp, 2017. Disponível em: http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/468/820. Acesso em: 15 jan. 2020.

DAISH, Alice. Transforming a museum to be data-driven using R. **Anais da 12ª. Conferência Internacional de Curadoria Digital** - Upstream, Downstream: embedding digital curation workflows for data science scholarship and society. 20 a 23 de fevereiro de 2017. Edimburgo. Disponível em: data-drive museums Zenodo (PDF). Acesso em: 09 set. 2018

DIAS, G. A.; VIEIRA, A. A. N. Big data: questões éticas e legais emergentes. **Ci. Inf**. v. 42, n.2, maio/ago. 2013. p. 174-184.

DUMBILL, E. What is Big Data? In: **O'Reilly Media Inc. Big Data Now**: current perspectives. O'Reilly Media: California. 2012. Disponível em: http://www.oreilly.com/data/free/fles/big-datanow-2012.pdf. Acesso em: 09 set. 2018

EMBRAPA. **Termo de uso**: Base de Imagens de Sintomas de Doenças de Plantas (PDDB). Disponível em: https://www.digipathos-rep.cnptia.embrapa.br/jspui/static/doc/PDDB-Termo_de_uso.pdf. Acesso em: 06 de mar. 2020

EMBRAPA. **Base de Imagens de Sintomas de Doenças de Plantas (PDDB)**. Disponível em: https://www.digipathos-rep.cnptia.embrapa.br/jspui/. Acesso em: 06 de mar. 2020.

FAGUNDES, P. B.; MACEDO, D. D. J.; FREUND, G. P. A produção científica sobre qualidade de dados em big data: um estudo na Base de dados Web of Science. **RDBCI**: Rev. Digit. Bibliotecon. Cienc. Inf. v. 16. n. 1, jan./abr. 2017. p. 194-21.

FAPESP. Anexo I: Política para Acesso Aberto às Publicações Resultantes de Auxílios e Bolsas FAPESP. Disponível em: http://www.fapesp.br/12592. Acesso em: 06 de mar. 2020.

FOX, P.; HENDLER, J. Changing the Equation on Scientific Data Visualization. **Science 331**, n. 705, 2011. Disponível em: http://data2discovery.org/dev/wpcontent/uploads/2013/05/Fox-andHendler_Visualization_Science-2011-Fox-705-8.pdf. Acesso em: 09 set. 2018

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big Data concepts, methods and analytcs. **International Journal of Information Management**. n. 35, p. 137-144, 2015. Disponível em: www. sciencedirect.com/science/article/pii/S0268401214001066. Acesso em: 05 jan. 2019

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6 ed. São Paulo: Atlas, 2008. 220 p.

GILL, T. Metadata and the Web. In: **Introduction to metadata**. Los Angeles, CA, USA: Getty Research Institute, 2008. p. 20-38.

GILLILAND, A. J. Setting the Stage. In: BACA, M. (Ed.) **Introduction to metadata.** Los Angeles: Getty Publications, 2016. Disponível em: http://www.getty.edu/publications/intrometadata. Acesso em: 10 mai. 2018

GODOY, Arilda Schmidt. Pesquisa qualitativa: tipos fundamentais. **Revista de Administração de Empresas**, São Paulo, v. 35, n. 3, p. 20-29. 1995

GOLDER, S.A.; HUBERMAN, B. A. **The structure of collaborative tagging systems**. 2005. Disponível em: < http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf>. Acesso em: 25 maio 2019.

GUEDES, Roger de Miranda. **A abordagem dialógica na indexação social**. 2010. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

GUEDES, Roger de Miranda; DIAS, Eduardo José Wense. Indexação social: abordagem conceitual. **Revista ACB:** Biblioteconomia em Santa Catarina, Florianópolis, v. 15, n. 1, p. 39-53,jan./jun. 2010. Disponível em: http://revista.acbsc.org.br/racb/article/view/686&gt. Acesso em: 20 mar. 2018

GUY, M; TONKIN, E. Folksonomies: tidying up tags? **D-Lib Magazine**, Reston, v. 12, n. 1, jan. 2006. Disponível em: http://www.dlib.org/dlib/january06/guy/01guy.html. Acesso em: 25 mar. 2019.

HAMBURGER, Amélia Império, *et al*. **FAPESP 40 anos abrindo fronteiras**. São Paulo: Editora da Universidade de São Paulo, 536 p. 2004.

HASSAN-MONTERO, Y. Indización social y recuperación de información. **No Solo Usabilidad Journal**, Granada, n. 5, nov. 2006. Disponível em: http://www.nosolousabilidad.com/articulos/indizacion_social.htm. Acesso em: 09 set. 2018

HJORLAND, B. **Is classification necessary after Google**? ISKO. 2012. p. 19-30.

HJORLAND, B. **Theories of knowledge Organization**. ISKO. 2013

HODSON, Simon; [et al.]. FAIR Data Action Plan: interim recommendations and actions from the European Commission Expert Group on FAIR data. 2018. June. 21 p. Disponível em: https://zenodo.org/record/1285290#.WyOW GS B9jt4. Acesso em: 30 jan. 2020.

HODSON, Simon; Jones, Sarah; Collins, Sandra; [et al.]. Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data. 2018. June. Disponível em: https://zenodo. org/record/1285272#.W7uasPmYSCg. Acesso em: 30 jan. 2020

IBEKWE-SANJUAN, F.; BOWKER, G. C. Implications of big data for knowledge organization. **Knowledge Organization**, v. 44, n. 3, p. 187-198. 2017

INSTITUTO Tecnológico de Aeronáutica. Plano de Desenvolvimento Institucional 2011 – 2020. Disponível em: http://www.ita.br/sites/default/files/pages/PDI_ITA%20Partes%201%2C%202%20e%203.pdf. Acesso em: 06 de mar. 2020.

INSTITUTO Tecnológico de Aeronáutica. Legislação do ITA. Disponível em: http://www.ita.br/adm/legislacao. Acesso em: 06 de mar. 2020.

JIN, Xiaolong, *et al*. Significance and challenges of big data research. **Big Data research 2**. p. 59 -64, 2015.

LANCASTER, F.W. **Indexação e resumos**: teoria e prática. 2.ed. Brasília: Briquet de Lemos, 2004

LAVILLE, Christina; DIONNE, Jean. **A construção do saber:** manual de metodologia de pesquisa em ciências humanas. Belo Horizonte: Artmed, 1999.

LYNCH, C. A. Big data: how do your data grow? **Nature**, v. 455, n. 7209, p. 28–29, 2008. Disponível em: http://www.nature.com/nature/journal/v455/n7209/full/455028a.html?foxtrotcallback=true. Acesso em: 30 jan. 2020.

LYNCH, C. How do your data grow?.***Nature* 455,** p. 28-29 2008. Disponível em: https://doi.org.ez31.periodicos.capes.gov.br/10.1038/455028a. Acesso em: 29 dez. 2019.

LYON, L. **Dealing with data**: roles, rights, responsibilities and relationships. Bath: UKOLN, 2007. Disponível em: https://purehost.bath.ac.uk/ws/portalfiles/portal/419529/dealing_with_data_report-final.pdf. Acesso em: 31 dez. 2019.

Manual de autodepósito de dados de pesquisa RI-UFSCar Versão v.1.1 2019

MASSONI, L. F. H; FLORES, A. B. A cidade representada em tags: explorando a folksonomia no Flickr. **PontodeAcesso**, Salvador, v. 11, n. 3, p. 133-147, dez. 2017

MATHES, A. **Folksonomies**: cooperative classification and communication through shared metadata. 2004. Disponível em: http://adammathes.com/academic/computer-mediated-communication/folksonomies.html. Acesso em: 09 set. 2018

MEDEIROS, C. B. FAPESP lança Rede de Repositórios de Dados Científicos do Estado de São Paulo. **Agência FAPESP**: 20 dez. 2019. Entrevista concedida a Maria Fernanda Ziegler. Disponível em: http://agencia.fapesp.br/fapesp-lanca-rede-de-repositorios-de-dados-cientificos-do-estado-de-sao-paulo/32251/. Acesso em: 20 dez. 2019

MELERO, Remedios. **Recomendaciones para la gestión de datos de investigación**: dirigidas a investigadores. 2018, 15 p. Disponível em: https://digital.csic.es/bitstream/10261/173801/1/Maredata-recomendaciones-ESP.pdf. Acesso em: 20 dez. 2019

MERHOLZ, P. **Metadata for the masses**. 2004. Disponível em: http://adaptivepath.com/ideas/e000361. Acesso em: 29 mar. 2018.

MILAGRE, J. A.; SANTAREM SEGUNDO, J. E. As contribuições da Ciência da Informação na perícia em informática no desafio envolvendo a análise de grandes volumes de dados- big data. **Informação & Tecnologia (ITEC)**. v. 2, n. 2, jul./dez. 2015, p. 35-48.

MILLER, E. An Introduction to the Resource Description Framework, D-Lib Magazine,
mai. 1998. Disponível em: <http://www.dlib.org/dlib/may98/miller/05miller.html>. Acesso em: 10 set. 2019.

MORVILLE, P. **Ambient findability**: what we find changes who we become. Cambridge: O'Reilly. 2005

MOSTAFA, S. P.; CRUZ, D. V. N.; AMORIM, I. S. Primavera nos dentes: fuga e resistência na era digital. **Liine em Revista**. v. 11, n. 2, nov. 2015, p. 360-374.

MOTTA, Fernanda Miranda de Vasconcellos; BARBOSA, Cátia Rodrigues; BARBOSA, Ricardo Rodrigues. Big data como fonte de inovação em museus: o estudo de caso do museu britânico. **Inf. & Soc.:Est**., João Pessoa, v. 29, n. 1, p. 83-100, jan./mar. 2019

MOURA, M. A. Folksonomias, redes sociais e a formação do tagging literacy: desafios para a organização da informação em ambientes colaborativos virtuais. **Inf. Inf.**, Londrina, v.14, n.esp., 2009.

NHACUONGUE, Januário Albino; FERNEDA, Edberto. O Campo da Ciência da Informação: contribuições, desafios e perspectivas. **Perspectivas em Ciência da Informação**, [S.l.], v. 20, n. 2, p. 3-18, jun. 2015. ISSN 19815344. Disponível em: http://portaldeperiodicos.eci.ufmg.br/…/arti…/view/1932/1591. Acesso em: 17 abr. 2018

NSF. National Science Foundation. Disponível em: https://www.nsf.gov/. Acesso em: 14 nov. 2019

OLIVEIRA, J. G. et al. Criação e ampliação de ontologias por folksonomias: uma revisão sistemática da literatura. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18, 2017, Marília. **Anais** [...]. Marília: UNESP.

ORGANISATION for Economic Co-Operation and Development. OECD principles and guidelines for access to research data from public funding. Paris; 2007. Disponível em: https://www.oecd.org/sti/sci-tech/38500813.pdf. Acesso em: Acesso em: 06 jan. 2020

PATO, Paulo Roberto Gomes. Imagens: polissemia versus indexação e recuperação da informação. 2015. 340 f., il. Tese (Doutorado em Ciência da Informação) Universidade de Brasília, Brasília, 2015.

PIERRO, Bruno de. Uma estratégia para dados. **Revista Pesquisa**, n. 267, São Paulo, p. 36-39, maio 2018. Disponível em: http://revistapesquisa.fapesp.br/wpcontent/uploads/2018/05/036-039_Gestao-de-Dados_267.pdf. Acesso em: 14 nov. 2019.

PIMENTA, R. M. Big data e controle da informação na era digital: tecnogênese de uma memória a serviço do mercado e do estado. **Tendências da pesquisa brasileira em Ciência da Informação**. v. 6, n. 2, jul./dez. 2013.

QUINTARELLI, E. Folksonomies: power to the people. In: INCONTRO ISKO ITALIA - UNIMIB, Milão, 2005. **Papers**... Milan: Universitá di Milano, 2005. Disponível em: http://www.iskoi.org/doc/folksonomies.htm. Acesso em: 27 mar. 2018.

RAUEN, Cristiane Vianna. A relevância de uma política nacional de acesso aberto a dados de pesquisa. **Revista Construção**, jan. 2018. Disponível em:<http://revistaconstrucao.org/ciencia-e-tecnologia/relevancia-de-uma-politicanacional-de-acesso-aberto-dados-de-pesquisa/>. Acesso em: 14 dez. 2020.

RIBEIRO, C. J. S. Big Data: os novos desafios para o profissional da Informação. **Informação & Tecnologia (ITEC)**: João Pessoa/Marília, v. 1, n. 1, p. 96-105, jan./jun., 2014

RODRIGUES, A. A.; DIAS, G. A. Big Data, algoritmos e a sociedade do controle. Encontro Nacional de Pesquisa em Ciência da Informação. v. 16, 2015.

RODRIGUES, Adriana Alves; DIAS, Guilherme Ataíde. Estudos sobre visualização de dados científicos no contexto da data Science e do Big Data. **Pesq. Bras. em Ci. da Inf. e Bib**., João Pessoa, v. 12, n. 1, p. 219-228, 2017.

SANTANA JUNIOR, C. A. et al. Uma ferramenta para recuperação de tags de blog baseado em microformatos. **Tendências da pesquisa brasileira em Ciência da Informação**. v. 7, n. 2, jul./dez. 2014.

SANTANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. **Inf. Inf**. v. 21, n. 2, maio/ago. 2016, p. 116-142.

SANTOS, R. F. dos; CORRÊA, R. F. Análise e síntese dos diversos usos do termo "Folksonomia" no âmbito da Ciência da Informação. In: ENCONTRO NACIONAL DA PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais eletrônicos** […] João Pessoa: UFPB, 2015a.

SANTOS, R.; CORRÊA, R. A folksonomia e a representação colaborativa da informação em ambientes digitais. **Tendências Da Pesquisa Brasileira Em Ciência Da Informação**, v. 8, n. 1. 2015b. Disponível em: http://ojs.lat/tpbci/index.php/tpbci/article/view/225. Acesso em: 09 set. 2018

SANTOS, R. F. dos. **Modelos colaborativos de indexação social e sua aplicabilidade na Base de Dados Referencial de Artigos e Periódicos em Ciência da Informação (BRAPCI)**. Recife, 2016. Dissertação (Mestrado) – Universidade Federal de Pernambuco, Centro de Artes e Comunicação, Ciência da Informação. 2016.

SAYÃO, L. F.; SALES, L. F. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, v. 21, n. 2, p. 90–115, 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939. Acesso em: 20 dez. 2019.

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, 2017. Disponível em: http://revistas.ufpr.br/atoz/article/view/49708. Acesso em: 20 set. 2019.

SAYÃO, Luís Fernando; SALES, Luana Farias. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. ***Tendências da Pesquisa Brasileira em Ciência da Informação***, 2013, vol. 6, n. 1.

SEMELER, Alexandre Ribas. **Ciência da Informação em contextos de e-science**: bibliotecários de dados em tempos de data science. Florianópolis, 2017. Tese (Doutorado) – Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós-Graduação em Ciência da Informação. 2017.

SHEARER, K. **Comprehensive brief on research data management policies**. [S.l: s.n.], 2015. Disponível em: http://docplayer.net/17594465-Comprehensivebrief-on-research-datamanagement-policies.html. Acesso em: 9 jan. 2020.

SILVA, J. L. C; FREIRE, G. H. de A. Um olhar sobre a origem da Ciência da Informação: indícios embrionários para sua caracterização identitária. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, v. 17, n. 33, p. 1-29, jan./abr., 2012. ISSN 1518-2924. DOI: 10.5007/1518-2924.2012v17n33p1

SILVA, J. V. da, SILVA, S. R. P. da. Gerenciamento do vocabulário de tags do usuário em sistemas baseados em folksonomia. **Assembla**, p. 201-204, 2008. Disponível em: https://dl.acm.org/purchase.cfm?id=1810040. Acesso em: 25 mar. 2019.

SMITH, G. **Atomiq: Folksonomy**: social classification. Aug 3, 2004. Disponível em: http://atomiq.org/archives/2004/08 /folksonomy_social_classification.html Acesso em: 29 mar. 2018.

SOUZA, R. R.; ALMEIDA, M. B.; BARACHO, R. M. A. Ciência da informação em transformação: big data, nuvens, redes sociais e web semântica. **Ci. Inf**. v. 42, n, 2, maio/ago. 2013, p. 159-173.

SPITERI, L. F. The structure and form of folksonomy tags: the road to the public library catalogue. In: BRAVO, B. R.; ALVITE DIEZ, M. L. eds. **La interdisciplinariedad y la transdisciplinariedad en la organización del conocimiento científico**. León: Universidad de León. 2007. p. 459-468

STONEBRAKER, M. **What does Big Data mean?** Communications of the ACM. 2012. Disponível em: http://cacm.acm.org/ blogs/blog-cacm/155468-what-does-big-data-mean/fulltext. Acesso em: 25 mar. 2019.

TARTAROTTI, R. D. E.; EVEDOVE, P. R.; FUJITA, M. S. L. Elaboração de tesauro conceitual de política de indexação para bibliotecas universitárias. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015. João Pessoa, **Anais...** João Pessoa: Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 2015.

TERENCE, Ana Cláudia Fernandes; ESCRIVÃO FILHO, Edmundo. Abordagem quantitativa, qualitativa e a utilização da pesquisa-ação nos estudos organizacionais. In:ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO, 26., 2006, Fortaleza. **Anais** [...], Fortaleza: ENEGEP, 2006.

TOMAÉL, M. I; SILVA, T. E. da. Repositórios institucionais: diretrizes para políticas de informação. In: Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB, 8., 2007, Salvador, **Anais...** . Salvador: UFBA, 2007. Disponível em: http://www.enancib.ppgci.ufba.br/artigos/GT5--142.pdf. Acesso em: 10 dez. 2019

TRANT, J. Studying social tagging and folksonomy: a review and framework. Toronto, Canada: **Journal of Digital Information**, 2008, Special Issue on Digital Libraries and User-Generated Content, 44 p.

UNIVERSITY OF MELBOURNE. Management of research data and records policy (MPF1242) [Internet]. Parkville; 2013 Nov. Disponível em:

https://policy.unimelb.edu.au/MPF1242#section-5. Acesso em: Acesso em: 06 jan. 2020

UNIVERSIDADE Federal do ABC. Plano de dados abertos. Disponível em: http://www.ufabc.edu.br/acesso-a-informacao/plano-de-dados-abertos. Acesso em: 06 de mar. 2020.

UNIVERSIDADE Federal do ABC. Repositório de dados de pesquisa UFABC. Disponível em: https://www.unifesp.br/reitoria/bibliotecas/images/bibliotecas/ufabc.pdf. Acesso em: 06 de mar. 2020

UNIVERSIDADE Federal do Estado de São Paulo. Repositório de dados de pesquisa UNIFESP. Disponível em: https://www.unifesp.br/reitoria/bibliotecas/images/bibliotecas/ufabc.pdf. Acesso em: 06 de mar. 2020

UNIVERSIDADE Estadual Paulista “Júlio de Mesquita Filho”. Repositório Institucional UNESP. Disponível em: https://repositorio.unesp.br/handle/11449/183294. Acesso em: 06 de mar. 2020

UNIVERSIDADE Estadual Paulista “Júlio de Mesquita Filho”. Política de Gestão do Repositório Institucional da Unesp Regulamento interno. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/144653/regulamento_repositorio_institucional_unesp.pdf?sequence=2&isAllowed=y. Acesso em: 06 de mar. 2020

UNIVERSIDADE Federal de São Carlos. Repositório Institucional UFSCar. Disponível em: https://repositorio.ufscar.br/. Acesso em: 06 de mar. 2020

UNIVERSIDADE Federal de São Carlos. Manual de autodepósito de dados de pesquisa RI-UFSCar. Disponível em: https://repositorio.ufscar.br/static/ri-ufscar-manual-autodeposito-dados-v1.1.pdf. Acesso em: 06 de mar. 2020

UNIVERSIDADE de Campinas. Repositório de Dados de Pesquisa da Unicamp. Disponível em: https://redu.sbu.unicamp.br:8181/dataverse/redu. Acesso em: 06 de mar. 2020

VANDER WAL, Thomas. **Folksonomy definition and Wikipedia**. 2007

VIANA, Joyce Mirella dos Anjos; ARAKAKI, Ana Carolina Simionato; DAL’EVEDOVE, Paula Regina. A contribuição das folksonomias para a construção de metadados. **VIII SECIN**. p. 509-523. Disponível em: http://www.uel.br/eventos/cinf/index.php/secin2019/secin2019/paper/viewFile/597/396. Acesso em: 06 jan. 2020.

VIANA, Joyce Mirella dos Anjos; DAL’EVEDOVE, Paula Regina; GRACIOSO, Luciana de Souza. Observações Pragmáticas na Indexação Social. In: BARROS, Thiago Henrique Bragato; TOGNOLI, Natalia Bolfarini (org.). **Organização do Conhecimento responsável: promovendo sociedades democráticas e inclusivas.** Belém: Ed. da UFPA, 2019. p. 77-84. Disponível em: http://isko-

brasil.org.br/wp-content/uploads/2019/09/LIVRO-ISKO-BRASIL-EDI%C3%87%C3%83O-BEL%C3%89M.pdf. Acesso em: 06 jan. 2020.

VIANA, Joyce Mirella dos Anjos; DAL'EVEDOVE, Paula Regina; TARTAROTTI, Roberta Cristina Dal'Evedove. A inserção da Folksonomia nos Programas de Pós-graduação em Ciência da Informação no Brasil. In: BARROS, Thiago Henrique Bragato; TOGNOLI, Natalia Bolfarini (org.). **Organização do Conhecimento responsável: promovendo sociedades democráticas e inclusivas.** Belém: Ed. da UFPA, 2019. p. 321-330. Disponível em: http://isko-brasil.org.br/wp-content/uploads/2019/09/LIVRO-ISKO-BRASIL-EDI%C3%87%C3%83O-BEL%C3%89M.pdf. Acesso em: 06 jan. 2020.

VOSS, J. Tagging, folksonomy e co-renaissance of manual indexing?. In: International Symposium for Information Science, 10., 2007, Cologne. **Anais** […].Cologne: Itália, 2007. p. 1-12. Disponível em: arxiv:cs/0701072v1. Acesso em: 29 mar. 2018.

WAMBA, Samuel Fosso *et al*. How big data can make big impact: Findings from a systematic review and a longitudinal case study. **International Journal of Production Economics,** v. 165, p. 234-246, 2015.

WALPORT, M.; BREST, P. Sharing research data to improve public health. **The Lancet**, v. 377, n. 9765, p. 537–539. 2011.

WILLIS, C.; GREENBERG, J.; WHITE, H. Analysis and synthesis of metadata goals for scientific data. **Journal of the American Society for Information Science and Technology**, v. 63, n. 8, p. 1505–1520, 2012. Disponível em: http://doi.wiley.com/10.1002/asi.22683. Acesso em: 8 set. 2019.