Biblos E-ISSN 2236-7594

AFINAL, O QUE É DADO DE PESQUISA?

Luis Fernando Sayão
Doutor em Ciência da Informação. Comissão Nacional de Energia Nuclear, Rio de Janeiro, Rio de Janeiro, Brasil.
lsayao@cnen.gov.br
http://orcid.org/0000-0002-6970-0553

Luana Sales
Doutora em Ciência da Informação. Instituto Brasileiro de Informação em Ciência e Tecnologia, Rio de Janeiro, Rio de Janeiro, Brasil.
luanafsales@gmail.com
https://orcid.org/0000-0002-3614-2356

RESUMO

No contexto científico contemporâneo em que a geração, coleta e análise intensiva de dados se tornam etapas essenciais para o progresso científico, os sistemas de gestão e curadoria de dados se reconfiguram como infraestruturas-chave na composição das ciberinfraestruturas de pesquisa. Porém, a curadoria e gestão de dados de pesquisa só se efetivam se forem consideradas as especificidades dos domínios disciplinares, os fluxos de pesquisa e, sobretudo, as características dos dados coletados, aproximando o ciclo de vida da pesquisa com o ciclo de vida dos dados. Com o objetivo de contribuir para a superação desse desafio, o presente estudo procura explicitar e discutir os principais conceitos, características e propriedades dos dados de pesquisa – sistematizando-os na forma de uma taxonomia – em função dos impactos que eles têm na concepção das infraestruturas tecno-socias e organizacionais dos sistemas de gestão e curadoria de dados de pesquisa.

Palavras-chave: Dados de Pesquisa. Curadoria de Dados. Conceituação.

AFTER ALL, WHAT IS RESEARCH DATA?

ABSTRACT

In the contemporary scientific context in which the generation, collection and intensive analysis of data become essential steps for scientific progress, data management and curation systems are reconfigured as key infrastructures in the composition of research cyber infrastructures. However, curation and data management is only effective if we consider the specificities of the disciplinary domains, the research flows and, above all, the characteristics of the collected data, bringing the research life cycle closer to the data life cycle. In order to contribute to overcoming this challenge, the present study seeks to explain and discuss the main concepts, characteristics and properties of research data – systematizing them in the form of a taxonomy – according to the impacts they have on the design of technological infrastructures. social and organizational aspects of research data management and curation systems.

Keywords: Research Data. Data Curation. Conceptualization.

Recebido em: 23/08/2020 Aceito em: 23/10/2020 Publicado em: 31/12/2020

1 INTRODUÇÃO

A palavra “dado” é o plural do latim *datum*, que significa dádiva, oferta ou algo reconhecido e usado como base para cálculos, é o vínculo primordial com os fenômenos de amplo espectro em que estamos imersos. O mais remoto método para registrar dados pode ter sido entalhes em bastões de madeira usados como forma de registrar a passagem dos dias, ou estacas fincadas no solo para marcar o nascer do sol no solstício,

porém há inúmeras outras histórias que marcam a ancestralidade do termo. Mais tarde, o ábaco foi inventado para ajudar o cálculo desses dados, e o desenvolvimento da escrita ampliou significativamente a capacidade humana de registrar experiências e eventos do nosso mundo, aumentando vertiginosamente a quantidade de dados coletados. Nos últimos 150 anos o desenvolvimento de sensores elétricos, a digitalização de dados e a invenção do computador contribuíram para um crescimento massivo da quantidade de dados que são coletados e armazenados (KELLEHER; TIERNEY, 2018, p. 6). Hoje em dia é difícil fugir da palavra “dado”, porém, como muitos outros conceitos que são cooptados para distintos propósitos, o termo “dado” tem diferentes significados que dependem fortemente do contexto onde estão sendo usados (SWANSON; RINEHART, 2016), mesmo quando nos referimos ao ambiente específico da pesquisa.

No mundo das ciências, muitas vezes associado, numa primeira visão, às ciências exatas e naturais, dados de pesquisa podem ser encontrados em qualquer disciplina e em muitas formas. Podem ser dados brutos coletados diretamente por um instrumento ou um sensor e agregados a partir de múltiplas fontes; ou podem ser produtos de um modelo teórico, simulação ou visualização; ou de experimentos conduzidos na bancada de um laboratório; ou ainda podem ser textos, bibliotecas de imagens digitais e modelos em 3D, tais como os usados para a reconstrução de sítios históricos e mitológicos. Os cientistas sociais produzem grande quantidade de dados, incluindo dados de levantamentos e observacionais, como por exemplo, as atividades e interações humanas complexas capturadas por sensores ou vídeos (JOHNSTON, 2017); o mesmo pode se dizer das pesquisas em saúde, que geram e analisam volumes massivos de dados, cuja importância vital para a sociedade tem como contrapartida rigorosas exigências em relação à gestão, que precisam levar em conta o alto grau de sensibilidade de suas informações.

Os dados têm uma outra face importante que se desenrola fora dos domínios da ciência, mas que, no entanto, são mundos cujas fronteiras se sobrepõem. Os negócios, a indústria e mesmo o governo reconfiguram a importância dos dados apoiados pelos desenvolvimentos vertiginosos das tecnologias digitais e pelo poder dos algoritmos, dos métodos de mineração de dados, aprendizagem de máquina e inteligência artificial que se amalgamam para produzir o fenômeno onipresente do *big data* e da ciência dos dados (ANDERSON, 2008; SAYÃO; SALES, 2019a).

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

No contexto científico em transição, em que geração e análise intensiva de dados definem novas metodologias imprescindíveis para o progresso das pesquisas, os sistemas de gestão e curadoria de dados de pesquisa se tornam infraestruturas essenciais ao lado de outros sistemas, dispositivos e equipamentos de apoio à pesquisa e fazem parte de uma configuração integradora chamada ciberinfraestrutura de pesquisa (GOLD, 2007). Todavia, uma boa e efetiva curadoria e gestão de dados somente se concretizará se for feita considerando as especificidades do domínio, os fluxos de pesquisa e, ainda, as características dos dados gerados nesse domínio de pesquisa. Isso significa dizer que estudar as propriedades do dado e como ele se manifesta em cada disciplina é condição necessária para a construção de critérios que tornarão o dado de pesquisa passível de ser selecionado, arquivado e preservado, de acordo com suas características. É preciso, portanto, aproximar o ciclo de vida da pesquisa ao ciclo de vida dos dados, ambos os contextos extremamente variáveis. Com o objetivo de contribuir para a superação desse desafio, o presente ensaio procura explicitar e discutir os principais conceitos, características e propriedades dos dados de pesquisa em função do impacto que eles têm na definição de infraestruturas tecno-sociais e organizacionais voltadas para a gestão desses ativos informacionais.

2 PROTAGONISMO DOS DADOS

Antes de tudo, é preciso considerar que a ubiquidade dos dados não é um fenômeno unicamente do nosso tempo. O governo, as empresas, a pesquisa científica, bem como vários outros segmentos da sociedade sempre lançaram mão de dados e informações para tomar decisões, redirecionar seus empreendimentos, fundamentar suas descobertas. Porém, nas últimas décadas, toda a sociedade experimenta um fenômeno inédito que tem como ponto de inflexão uma mudança na curva de disponibilidade de informação: da escassez à extrema abundância de dados. Isso muda muita coisa no mundo em que vivemos, é uma revolução que está transformando o modo como vivemos, trabalhamos, nos divertimos e como produzimos conhecimento científico (SAYÃO; SALES, 2019a).

“Dados e informações sempre foram *input* e *output* da pesquisa científica. O que é novo é a escala de dados e informação envolvidos” (BORGMAN, 2007 p.6). O dilúvio de dados, já há muito previsto pela comunidade científica, chegou às esferas da ciência transformando seus métodos seculares de construção de novos conhecimentos.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

Sistemas avançados de computadores e de redes permitem que conjuntos massivos de dados sejam integrados e explorados de forma que revelem relações e padrões intrínsecos, porém insuspeitos a “olho nu”. Esta ciência orientada por dados é uma promissora fonte de novos conhecimentos (THE ROYAL SOCIETY, 2012). De fato, há grandes expectativas em torno de um mundo rico em dados que inclui a descoberta de novas drogas, uma compreensão mais precisa das mudanças climáticas e o aprimoramento da capacidade dos pesquisadores examinarem a história, a cultura e a arte (BORGMAN, 2012, p.2). Porém, ao mesmo tempo que essas mudanças desencadeadas pelo uso intensivo de dados digitais impactam de forma contundente os processos de pesquisa, também modificam nosso entendimento sobre as infraestruturas de informações para a pesquisa necessárias a esse novo paradigma científico que se delineia – o chamado quarto paradigma.

O crescimento de dados produzidos e consumidos pelas chamadas “*big sciences*”, tais como física e astronomia, redefine um novo modelo de ciência conhecido coletivamente como “quarto paradigma” e o surgimento de campos híbridos de estudo, tais como astroinformática, biologia computacional e humanidades digitais (BORGMAN, 2012). Nesse contexto, o volume de dados científicos que são gerados por projetos de pesquisa altamente instrumentalizados – como aceleradores lineares, redes de sensores remotos, sismógrafos, entre muitos outros – é tão robusto que só pode ser capturado e gerenciado por meio do uso intenso de tecnologias de informação, de redes de distribuição por computação em grade e por novas concepções de colaboração e compartilhamento. Isto porque a quantidade de dados produzidos excede de muito a capacidade de gestão por métodos e sistemas tradicionais. Mas esse fenômeno não está limitado aos domínios da grande ciência: a cauda longa da ciência – formada coletivamente por inúmeros projetos de pesquisa desenvolvidos em pequenos laboratórios – está se tornando cada vez mais intensiva na geração e análise de dados, na medida em que novos métodos e instrumentação permitem que os investigadores individuais e pequenas equipes coletem volumes sem precedentes de dados; por seu lado, os cientistas sociais estão analisando um volume cada vez maior de dados provenientes das estatística governamentais, levantamentos *on-line* e modelos comportamentais; da mesma forma, os acadêmicos humanistas estão produzindo e analisando uma grande quantidade de textos, imagens e vídeos digitais, dados de redes sociais e modelos de sítios históricos.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

Face a essas mudanças, os veículos de comunicação científica estão também se transformando: dados, modelos e visualização são incorporados aos artigos de periódicos e de conferência e a outros produtos de pesquisa. Documentos acadêmicos se tornam objetos digitais complexos, como são as publicações ampliadas e os *datajournals,* e se reconfiguram em conceitos inéditos de publicação, como os periódicos de resultados negativos (SALES; SAYÃO 2019b). Novas ferramentas e serviços são desenvolvidos para produzir, publicar, contextualizar, distribuir e gerenciar esses novos conceitos de mídia acadêmica. Tudo indica que os padrões e fluxos da comunicação científica estão rapidamente se ajustando às novas normas de publicação e compartilhamento de dados de pesquisa (BORGMAN, 2007). Subjacente a esses desenvolvimentos, as tecnologias de comunicação e de rede maximizam o potencial de criar novas dinâmicas sociais em prol do compartilhamento de dados de pesquisa. Concepções inéditas de reconfiguração dos dados, como a tecnologia de dados ligados, geram novas informações por meio de profunda integração de dados provenientes de diferentes conjuntos de dados, que aumentam as possibilidades de análises automáticas (THE ROYAL SOCIETY, 2012). Dados digitais são ubíquos e rapidamente remodelam como a pesquisa acadêmica progride agora e no futuro. A abundância – algumas vezes caótica – do fluxo planetário de dados engendra novas formas de exploração colaborativa e de descoberta que minimiza as barreiras internacionais e interdisciplinares, conectando pesquisadores com objetivos compartilhados e acelerando as taxas de entendimento científico.

Essa abundância de dados de pesquisa em formatos digitais, todavia, desafia as bibliotecas e os profissionais da ciência da informação a explorar o potencial desses fluxos de informações que partem das descobertas de pesquisas, dos laboratórios e atividades acadêmicas e a preservar as evidências únicas para uso futuro. Existem muitas motivações para armazenar dados, porém o objetivo mais importante é o reuso por outros pesquisadores (JOHNSTON, 2017), agora e no futuro. Mas para isso, é necessária uma compreensão sobre o que é dado nos diversos contextos científicos que ele aparece. Portanto, a questão que se coloca é “o que são dados?”, porém muito frequentemente a questão se torna “quando são dados?”, posto que alguns fenômenos que podem ser tratados como dados são, por si só, atos acadêmicos; porém fatos, registros e situações cotidianas podem, também, se transformar em *inputs* para a pesquisa, como uma carta, uma fotografia ou um coletivo de mensagem numa rede social. “As letras iniciais

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

iluminadas de manuscritos medievais destinavam-se a ser decorativas, mas se tornaram uma fonte importante sobre roupas medievais e utensílios" (BUCKLAND, 1999, p.355).

3 DEFINIÇÕES E CONTEXTOS

O que dificulta atribuir uma definição consensual ao dado de pesquisa é o fato idiossincrático de que ele pode ser muitas coisas diferentes para pessoas e circunstâncias diferentes. Isto acontece porque o dado de pesquisa é dependente de interpretação. "Informação é um conceito complexo com centenas de definições. Dado é um conceito simples com poucas definições, mas sujeito a muitas e diferentes interpretações" (BORGMAN 2007, p.119). Com essa contraposição, que enfatiza o potencial de ressignificação dos dados, Christine Borgman propõe que dado seja "uma representação passível de reinterpretação que se apresenta de uma maneira formalizada adequada para comunicação, interpretação ou processamento" (2007, p.119), como preconiza o Modelo de Referência OAIS (CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM, 2012, p.1-10).

De fato, o termo "dado de pesquisa" 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. Pode ser o resultado de um experimento realizado num ambiente controlado de laboratório, um estudo empírico na área de ciências sociais ou a observação de um fenômeno cultural ou da erupção de um vulcão num determinado momento e lugar. Dados digitais de pesquisa ocorrem na forma de diferentes tipos de dados, como números, figuras, vídeos, softwares; com diferentes níveis de agregação e de processamento, como dados crus ou primários, dados intermediários e dados processados e integrados; e em diferentes formatos de arquivos e mídias. Essa diversidade que vai sendo delineada pelas especificidades de cada disciplina, suas condicionantes metodológicas, protocolos *workflows* e seus objetivos, se torna um desafio, mesmo para o pesquisador, pelo alto grau de contextualização necessário, definir precisamente o que é dado de pesquisa de uma forma transversal aos diversos domínios disciplinares (BORGMAN, 2010; PAMPEL *et al*., 2013).

"Dados são sempre registrados tomando como base de algum interesse, perspectiva, tecnologia e prática que determinam seus significados e utilidades em diferentes contextos" (NIELSEN, HJØRLAND, 2014, p.225).Esta citação ilustra bem o fato

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

de que o significado de dados é determinado por uma forte contextualização de diferentes níveis, escalas e granularidade que inviabiliza que tenhamos uma definição de “tamanho único” para toda a ciência; além do mais, é um termo em evolução sincronizada com as amplas tendências sociais e técnicas e pelo seu uso intenso e valor ascendente para diversas finalidades.

O termo dado é usado, algumas vezes, como sinônimo de informação, enquanto em outras situações se refere unicamente a números. A inexatidão do termo e sobre o que ele se refere constituem um desafio em compreender devidamente as rápidas mudanças na área de dados (SWANSON; RINEHART, 2016). Por outro lado, as definições encontradas nos dicionários e enciclopédias falham em capturar a riqueza e a variedade dos dados no mundo da ciência ou em revelar as premissas epistemológicas e ontológicas sobre as quais eles são baseados (BORGMAN, 2015). Na esfera acadêmica, grande parte das definições são uma enumeração de exemplos como a oferecida – e bastante utilizada – pelo National Research Council (1999), que informa que dados são fatos, números, letras e símbolos. Contudo, “listas de exemplos não são verdadeiramente definições, posto que não estabelecem uma clara fronteira entre o que inclui e o que não inclui o conceito” BORGMAN (2015, p.19); porém, podem, na melhor das hipóteses, ser um ponto de partida para a compreensão do que pode ser dado, para algum propósito, em algum ponto do tempo, conclui a autora.

No seu mais profundo cerne, “dado pode ser qualquer informação que é factual e que pode ser analisada” (JOHNSTON, 2017, p.2). Essa definição minimalista caracteriza o coletivo de definições que se apresentam, que, via de regra, são parcimoniosas e procuram contornar a complexidade estrutural e semântica do conceito de dado e clamam por uma robusta contextualização. Por exemplo, para o Digital Curation Centre (DCC), no âmbito específico do seu Modelo do Ciclo de Vida da Curadoria que envolve dados, objetos digitais e bases de dados sobre camadas de ações de curadoria, dado é “qualquer informação em forma digital binária”. Esta definição é intencionalmente muito ampla e estende-se além da estreita conexão do mundo com os resultados da pesquisa científica (HARVEY, 2010).

Na esfera política, administrativa ou organizacional, a visão sobre os dados tem outras acepções. Por exemplo, a Organização para Cooperação e Desenvolvimento Econômico (OECD), em seu guia para acesso aos dados de pesquisas financiadas

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

por recursos públicos, define como dados de pesquisa “registros de fatos usados como fontes primárias na investigação científica e que geralmente são aceitos na comunidade científica como necessários para a validação dos resultados da pesquisa” (OECD, 2007).

No contexto operacional, as definições deveriam se voltar para delinear os contornos de atuação dos sistemas de gestão e curadoria, revelando com clareza o que é dado de pesquisa no seu contexto de atuação. Isto porque as instituições que custodiam os dados precisam explicitar com que entidades elas lidam e de que forma devem fazê-lo.

No presente trabalho, que tem como objetivo estabelecer uma base terminológica funcional para dados de pesquisa, no sentido de subsidiar os contornos de sistemas de gestão e curadoria, considera-se os seguintes aspectos conceituais:

- **Aspectos lógicos – em que discute** o que é o objeto (tipo de registro, coletado, criado, observado ou utilizado no âmbito da pesquisa científica), definindo o que é o dado de pesquisa.
- **Aspectos ontológicos – em que discute** a relação do objeto com o mundo, ou seja, a relação do dado de pesquisa no ambiente científico, considerando definições que discutem sua função, aplicação e objetivos, ou dito de outra forma, como os dados de pesquisa podem ser interpretados, tratados e aceitos pela comunidade científica como evidência para analisar, validar e produzir resultados de pesquisa.

O que nos conduz à seguinte definição: dado de pesquisa é todo e qualquer tipo de registro coletado, observado, gerado ou utilizado no âmbito da pesquisa científica, que pode ser interpretado, tratado e aceito como evidência pela comunidade científica e necessário para analisar, validar e produzir resultados de pesquisa.

Essa definição é suficientemente ampla para abarcar todas as possibilidades de dados de pesquisa. No entanto, é importante destacar que para que o registro se configure como dado de pesquisa, ele precisa ser tratado e aceito pela comunidade. Isso coloca em pauta a obrigação de uma gestão mínima, com atribuição de metadados que tornem o registro compreensível para a comunidade científica.

Pela imprecisão das definições, circulares e nebulosas quanto aos contornos de seus objetos conceituais, parece mais provável que a gestão e a curadoria precisem se apoiar numa categorização dos dados mais consensual e rigorosa que possa subsidiar a criação de uma taxonomia voltada para apoiar as ações de gestão. É o que veremos a seguir.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

4 TIPOLOGIA DOS DADOS OU O QUE EXATAMENTE QUEREMOS CURAR?

Dados de pesquisa podem ser descritos de muitas formas, refletindo o alto grau de diversidade e as várias faces das atividades científicas. Os dados numa coleção podem incluir números, imagens, *streaming* de vídeo ou áudio, software e informações de versionamento de software, entrevistas, algoritmos, equações, animações ou modelo e simulações, numa variedade virtualmente infindável, quer requer, para a sua gestão e curadoria algum nível de categorização, pois precisamos saber "o que exatamente queremos curar" (HARVEY, 2010, p.45).

A tipificação e a caracterização de dados de pesquisa têm dois principais objetivos no contexto da curadoria: o primeiro deles é compreender com mais precisão a natureza dos dados, como eles se apresentam e seus fluxos de geração, processamento e uso; o segundo objetivo é, a partir do conhecimento dessa natureza complexa, construir modelos e estratégias, dimensionar infraestruturas tecnológicas e organizacionais para a sua gestão, além de determinar o grau de investimento na preservação de longo prazo. A heterogeneidade intrínseca aos dados de pesquisa implica também necessidade de formular políticas e estratégias de gestão de amplo espectro que englobem os vários tipos de dados coletados e analisados por uma instituição ou comunidade acadêmica. O reconhecimento dessas diferenças torna-se crucial para diversas ações no escopo da gestão de dados de pesquisa e do ciclo de vida da curadoria, posto que cada tipo de dado vai necessitar de processos distintos de gestão e de profundidade de ações de curadoria, como por exemplo, arquivamento de longo prazo. Caminhando nessa direção, a seguir analisaremos as características mais relevantes dos dados de pesquisa apontadas pela literatura.

4.1 A Origem dos Dados de Pesquisa

Dados de pesquisa podem ser distinguidos pela sua origem que nos remete aos métodos e processos através dos quais eles foram coletados ou criados. Há um consenso claro de que os dados científicos podem ser categorizados quanto à origem em **observacionais**, **computacionais** e **experimentais.**

Este princípio de divisão é, provavelmente, o mais crítico para o tratamento dos dados de pesquisa, pois estabelece a profundidade e a abrangência temporal da gestão

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

e curadoria que se deve aplicar para cada categoria em termos de arquivamento e preservação. Isto porque a diferença de criticidade conferida pela possibilidade – ou impossibilidade – de replicação desses dados implicarão adoção de estratégias distintas de gestão e curadoria. Dessa forma, a origem dos dados de pesquisa pode influenciar nas decisões operacionais sobre que coleções de dados merecem ser preservadas e por quanto tempo elas devem ser curadas (NATIONAL SCIENCE BOARD, 2005).

DADOS OBSERVACIONAIS – são originados de observações de fenômenos e eventos únicos, como a observação da temperatura do oceano numa data específica, da atitude de eleitores antes de uma eleição ou da fotografia de uma supernova. Dados observacionais são caracteristicamente únicos, não repetitíveis – ou seja, originados em eventos que não se repetirão jamais – e são, geralmente, coletados em grandes volumes (NATIONAL ARCHIVES AND RECORDS ADMINISTRATION, 2007). Portanto, são dados que constituem registros históricos que não podem ser capturados uma segunda vez, e, portanto, devem ser arquivados para sempre, com níveis necessários de fidedignidade, autenticidade e integridade, em estruturas confiáveis que garantam sua preservação por longo prazo (NATIONAL SCIENCE BOARD, 2005). Dados observacionais brutos são tipicamente coletados por meio de percepção humana ou mensuração (por exemplo, notas de campo) ou, mais comumente, por sensores ou por outros instrumentos. Borgman (2012) assinala que nas ciências exatas e biociências exemplos dessa categoria incluem observações de condições meteorológicas, de plantas e animais. Essas observações podem ser realizadas por satélite, rede de sensores ou ainda por caneta e caderno de notas; nas ciências sociais, os exemplos incluem indicadores econômicos, entrevistas e etnografias. A coleta de dados observacionais tem duas modalidades: ***in situ,*** quando a coleta é operacionalizada em contato direto com o fenômeno; **em contraste**, coleta à distância que envolve uso de instrumentos, tais como câmeras ou sensores remotos digitais, que não têm contato direto com os fenômenos estudados. A necessidade de curadoria de longo prazo para os dados observacionais decorre do fato deles estabelecerem um patamar que ajuda a determinar as taxas futuras de mudança e frequência de ocorrências de eventos não usuais, por exemplo, os parâmetros temporais do aparecimento de uma supernova. As coleções de dados que cobrem longos períodos permitem que padrões mais consistentes possam ser identificados e dessa forma aumentam a confiança na veracidade dos dados e das conclusões extraídas a partir

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

deles. “Além do mais, dados observacionais frequentemente podem ser processados e usados de formas inéditas, por exemplo, para verificar novos conceitos científicos (NATIONAL ARCHIVES AND RECORDS ADMINISTRATION, 2007, p.10)

DADOSEXPERIMENTAIS – são dados provenientes de procedimentos controlados em bancadas de laboratórios para testar ou estabelecer hipótese ou para descobrir ou testar novas leis. Exemplos incluem resultados de pesquisa em química, em laboratório de via úmida, experimentos de física em colisores lineares e experimentos em psicologia controlados em um laboratório ou que desenrolam em campo. Em tese, estes dados são provenientes de experimentos que podem ser precisamente reproduzidos e não precisam ser armazenados indefinidamente (Borgman, 2012). Entretanto, nem sempre é possível reproduzir precisamente todas as condições experimentais, especialmente quando algumas condições e variáveis experimentais podem não ser conhecidos; e quando o custo de reproduzir a pesquisa é proibitivo. Na maioria das vezes, o custo de gerenciar os dados é apenas uma fração do custo de refazer o experimento (THE ROYAL SOCIETY, 2012), o que viabiliza a preservação de longo prazo. Dessa forma, considerações sobre o custo e reprodutibilidade são relevantes quando se estabelecem políticas para a preservação de dados experimentais (NATIONAL SCIENCE BOARD, 2005).

DADOS COMPUTACIONAIS – são produtos de processamento de modelos computacionais, simulações e *workflow*. Embora bastante comum nas áreas de física e de biociências, essa categoria também é encontrada nas ciências sociais e humanidades. Dados computacionais geralmente não precisam de preservação de longo prazo em repositório ou centros de dado, posto que os programas podem ser executados novamente e os dados reproduzidos. Porém é necessário que esteja disponível uma grande quantidade de informações sobre o modelo – incluindo uma completa descrição do *hardware* e do *software* e dados de entrada, especialmente na forma de metadados. “Algumas vezes, a cópia do código usado para criar ou processar os dados é tão imprescindível para o uso dos dados que o código deve ser quase que pensado como parte dos ‘metadados’ que descrevem os dados” (GOODMAN *et al,* 2014, p.1). Assim, as saídas de um modelo não precisam ser preservadas, porém é essencial o arquivamento do modelo propriamente dito e de um conjunto robusto de metadados. Os dados computacionais estão relacionados ao advento da simulação computacional complexa, onde a modelagem matemática de fenômenos – que há muito tempo tem sido

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

uma ferramenta importante da ciência – cria uma articulação quantitativa formal a partir de uma teoria científica, permitindo predições e *insights*. É preciso observar que uma simulação computacional é análoga a um experimento físico, todavia é um experimento conduzido por meio de equações matemáticas ao invés de entidades físicas. Seus produtos, entretanto, podem ser considerados dados que são equivalentes aos dados produzidos por experimentos físicos. “Em princípio, portanto, não existe razão para que os dados provenientes de simulação devam ser considerados diferentes de outras formas de dados científico” (THE ROYAL SOCIETY, 2012, p.35). Entretanto, o poder de uma simulação ou sua capacidade de representar uma realidade depende de vários fatores, como: da acurácia da aproximação a uma formulação teórica do problema; do algoritmo numérico usado para resolvê-lo; e da precisão numérica do computador. Os modelos propriamente ditos, em algumas áreas como ciências biológicas, são considerados dados de pesquisa e podem estar disponíveis em repositórios, como por exemplo o BioModels[1].

Nas categorias quanto à origem, os dados observacionais, experimentais e computacionais são criados ou coletados em ambientes, contextos ou situações condicionadas ou controladas por rígidos protocolos científicos, como laboratórios, observatórios, expedições, ambientes científicos virtualizados e outros. São dados de pesquisa na sua essência. Entretanto, os dados governamentais, registros médicos – como os prontuários – e arquivísticos, dados provenientes de censo e dados extraídos das redes sociais, por exemplo, são dados coletados para outros propósitos que são frequentemente utilizados para a pesquisa científica. São dados usados para a pesquisa que têm sua origem fora do mundo científico. Borgman (2015) caracteriza esses dados como uma quarta categoria em relação à origem, chamada de **registros**, que abrangem formas de dados que não se enquadram facilmente nas categorias observacionais, experimentais ou computacionais ou que não resultam de nenhuma dessas categorias. Nessa perspectiva, “Registros de quase todos os fenômenos ou de qualquer atividade humana podem ser tratados como dados para a pesquisa”, conclui Borgman (2015, p.24).

É preciso notar que as fronteiras entre as categorias enunciadas acima são permeáveis, permitindo que haja uma profícua articulação transversal entre os diversos tipos de dados. Por exemplo, dados observacionais podem ser usados como *input* para experimentos e para refinar modelos computacionais; achados obtidos por meio

[1] Disponível em: https://www.ebi.ac.uk/biomodels.Acesso em: 21 ago. 2020

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

de experimentos e modelos são usados para aperfeiçoar métodos de observações, numa interlocução virtualmente ilimitada (NATIONAL SCIENCE BOARD, 2005).

A gestão do ciclo de vida dos dados de pesquisa implica, portanto, que os dados sejam avaliados por parâmetros que indiquem se eles serão ou não arquivados e por quanto tempo; como eles serão tratados e para que audiência eles se destinam – a comunidade-alvo – , que pode ser um determinado grupo de pesquisa, usuários de um repositório disciplinar ou para o reuso de um público não-especialista ou para todos eles. Replicações experimentais por outros pesquisadores, exigem especificações precisas sobre o processo inicial de aquisição de dados, manipulação e armazenamento; por outro lado, a replicação de dados obtidos por simulação pode exigir a especificação exata do ambiente computacional e o *dowloading* de máquinas virtuais, como o Virtual Observatory[2] (THE ROYAL SOCIETY, 2012). Os avanços tecnológicos vão mudando os padrões de gestão dos dados, por exemplo, em áreas como genômica, o custo de obter informação em sequenciamento genético está caindo mais que o custo de armazená-la, indicando que rapidamente deverá ser mais barato sequenciar novamente as amostras, quando for necessário, do que armazenar os dados. Porém, via de regra, os custos de gestão, correspondem a uma fração do custo de se gerar os dados novamente.

4.2 Quanto ao Grau de Processamento

Raramente os dados de pesquisa são utilizados logo após serem coletados pelos instrumentos científicos, eles demandam graus distintos de processamento, que refletem os fluxos de pesquisa de cada domínio disciplinar. Assim, as atividades de processamento e de curadoria geram o que chamamos de **dados derivados**, que são originados por ações sobre os **dados brutos**. Inicialmente, os dados são coletados na forma bruta ou primária, por exemplo, um sinal gerado por um instrumento ou sensor, como um pluviômetro; após coletar os dados brutos, os cientistas processam esses dados em diferentes níveis de complexidade. Assim, os dados brutos são frequentemente sujeitos a estágios subsequentes de refinamento, limpeza e análise, que geram diferentes versões, que estão mais proximamente relacionados com os objetivos da pesquisa. Cada nível de processamento adiciona valor aos dados por meio de ações, como resumir e interpretar os dados brutos e sintetizar novos dados. A produção de dados processados pode envolver

[2] Disponível em:http://www.ivoa.net/.Acesso em: 21 de ago. 2020

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

a incorporação de duas ou mais fontes – brutas ou processadas – para gerar outros produtos de dados. Por exemplo, a combinação de dados observacionais e dados de elevação para produzir uma representação tridimensional (NATIONAL ARCHIVES AND RECORDS ADMINISTRATION, 2007; NATIONAL SCIENCE BOARD, 2005).

Enquanto os dados brutos podem ser a forma mais completa de informação, pois sintetizam todas as informações captadas pelo instrumento, os dados derivados podem ser mais prontamente utilizados por outros pesquisadores na medida em que explicitam mais claramente determinadas facetas. De fato, os dados brutos, ou submetidos a baixos níveis de processamento, são mais difíceis de serem compreendidos e (re)usados por outros pesquisadores que não sejam os seus próprios criadores, posto que portam um componente de conhecimento tácito e, em muitos casos, são pouco documentados. Porém, os dados brutos são essenciais para se conduzir reanálises, tais como verificar achados ou apoiar novas hipóteses. Dessa forma, para facilitar futuras reanálises, é mais apropriado manter os dados de pesquisa nos níveis mais baixos de processamento compatíveis com o uso efetivo que se pretende. Contudo, se o objetivo for um reuso mais orientado, um alto nível de processamento torna os dados mais fáceis de serem usados por uma ampla audiência. É importante notar que, a maioria dos dados de pesquisa que estão sob a ação de curadoria em repositórios ou centros de dados, são dados processados e não dados brutos (THE ROYAL SOCIETY, 2012). Nesse contexto, é mais provável que os dados processados tenham mais valor de longo prazo quando a sua recriação for custosa ou difícil de ser realizada a partir dos dados brutos. Neste caso, pode ser importante arquivar de forma permanente tanto uma versão bruta quanto uma ou mais versões processadas de algumas coleções de dados.

O processo experimental estabelece outra distinção importante para os processos de gestão que têm como origem as diferenças entre os **dados intermediários** coletados durante as pesquisas preliminares e os **dados finais**. Os pesquisadores, muitas vezes, podem conduzir variações de um experimento ou coletar dados sob uma variedade de circunstâncias e relatam apenas os resultados que eles acham que são mais interessantes. Os dados finais selecionados são rotineiramente incluídos na coleção de dados; entretanto, os dados intermediários muito frequentemente não são arquivados ou permanecem inacessíveis para outros pesquisadores. Há, no entanto, a crescente percepção de que os dados podem ser úteis para outros pesquisadores. E isso dá origem

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

a uma reavaliação sobre o custo-benefício de se arquivar os dados intermediários. Em muitas situações aparece o conceito de dados terciários, que são os dados altamente condensados que são publicados em formas de tabelas, gráficos, diagramas, como parte de publicações acadêmicas, como um artigo.

Em decorrência dos níveis de processamento, é possível definir também níveis de uso dos dados. O **uso primário** de dados observacionais é frequentemente feito por pesquisadores envolvidos na coleta e no processamento de dados. **Usuários secundários** incluem cientistas e não cientistas. Pesquisadores geralmente exploram os dados de novas maneiras, dessa forma, podem criar dados adicionais processados. Por exemplo, fazendeiros usam dados de climatologia para decidir sobre a seleção de cultivo e engenheiros usam dados sísmicos para projetar estruturas críticas, como são as usinas nucleares. Dessa forma, o processamento e uso dos dados podem gerar outros produtos que não são dados, por exemplo: previsões climáticas e avisos de furacões, a partir de dados meteorológicos; carta de navegação e mapas elaborados, a partir de dados oceanográficos (NATIONAL ARCHIVES AND RECORDS ADMINISTRATION, 2007). Tomando em conta que muitos serviços são gerados por execução de modelos matemáticos, desse modo, no âmbito da curadoria de dados, a computação é tão importante na simulação e no processamento de dados observacionais e experimentais que frequentemente se torna difícil traçar uma linha que separe ‘dado’ de ‘análise’ (ou ‘código’), completa Goodman e seus colaboradores (2014).

4.3 Outras Características Relevantes

Muitas outras características dos dados de pesquisa são importantes para orientar uma gestão e curadoria apropriada. A seguir são incluídas algumas categorias baseadas em Sales e Sayão (2019c), que estão presentes na taxonomia proposta.

QUANTO À ABORDAGEM DA PESQUISA – a importância dessa categoria está alinhada com o tratamento distinto que os dados científicos precisam receber de acordo com a metodologia adotada no desenvolvimento da pesquisa. Uma parcela relevante dos dados de pesquisa é proveniente de estudos qualitativos originados nas áreas das ciências sociais, humanidades, arte e cultura, e mesmo das áreas das ciências exatas, que exigem infraestruturas tecnológicas e práticas distintas dos dados qualitativos.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

QUANTO À NATUREZA DOS DADOS – retrata a grande diversidade e heterogeneidade de tipos de dados de pesquisa que podem ser originados no ambiente de pesquisa em termos de formatos, mídias, suportes, expressões, arcabouço tecnológico e outros.

QUANTO AO NÍVEL DE SENSIBILIDADE – essa categoria explicita os níveis de abertura e compartilhamento durante o processo de gestão e curadoria, e a profundidade de intervenção que deve ser imposta às coleções, especialmente no caso das pesquisas que envolvem pessoas e informações pessoais sensíveis e informações confidenciais. Isto porque, nem todos os dados de pesquisa devem ser compartilhados sem tratamentos adequados, por exemplo, a submissão dos dados a processos de anonimização e armazenamento em sistemas *off-line.* Existem dados sensíveis, de grande utilidade para a ciência, que só poderão ser acessados por meios de condições controladas.

QUANTO À MATERIALIDADE – permite vislumbrar a necessidade de atribuir tratamento especial aos dados de pesquisa de acordo com a forma em que se materializam – digitais ou físicos. O tratamento proporcionado aos dados de pesquisa nascidos digitais, ou que passaram por processos de digitalização, tem características específicas determinadas pela fragilidade intrínseca dos objetos digitais e pela obsolescência tecnológica que ameaçam a sua capacidade de renderização, de autenticidade e de integridade. Por sua vez, as amostras físicas representam uma ampla variedade de dados de pesquisa. Elas são elementos básicos para referência, estudos e experimentação na pesquisa científica. Testes e análises são realizados diretamente baseados em amostras, tais como espécimes biológica, rochas e minerais, solos, sedimentos, plantas e sementes, artefatos arqueológicos, amostras de tecidos humanos e DNA; outros objetos físicos, como mapas e fotografias impressas são objetos diretos de estudos (RESEARCH DATA ALLIANCE, 2018). Cada um desses tipos de dados necessita de processos de curadoria e cuidados completamente diferentes, por exemplo, um herbário necessita de adubos, irrigação e níveis de temperaturas, umidade e insolação controlados. Porém, a sua forma de representação para fins de recuperação continua se valendo de metadados. Neste caso, em bases de dados referenciais.

QUANTO À PERENIDADE – as coleções de dados de pesquisa são gerenciadas e preservadas de acordo com a sua relevância para a ciência. No mais alto grau estão as coleções internacionais que são referências importantes que precisam de estruturas

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

estáveis que as mantenham para sempre, como o GenBank[3] e o Protein Data Bank[4]; em um nível intermediário estão as coleções nacionais e regionais; por fim, as coleções institucionais, comunitárias e individuais que têm valores mais transitórios que, porém, podem migrar para níveis mais perenes.

QUANTO À ABERTURA – diferentes níveis de abertura são atribuídos às coleções de dados pelos criadores e pelos sistemas de gestão. Esses níveis estão relacionados com o grau de sensibilidade dos dados, propriedade intelectual, interesses comerciais, patentes, segurança nacional, interesse pessoal do pesquisador e outros. Às coleções de dados de pesquisa também são atribuídas licenças que determinam o grau de reuso que terceiros podem fazer com essas coleções.

QUANTO À ESTRUTURAÇÃO – os dados de pesquisa podem ser estruturados, isto é, aqueles que são organizados segundo modelos estruturalmente bem definidos, como bases de dados relacionais, planilhas ou tabelas, por exemplo, dados demográficos levantados no censo; e dados de pesquisa não-estruturados – aqueles que podem ter sua própria estrutura interna no contexto de uma coleção, por exemplo, uma coleção de página web em que cada página tem uma estrutura diferente; outros exemplos: e-mails, tweets, texto de mensagens, músicas, vídeos e arquivos multimídia (KELLERHER; TIERNEY, 2018).

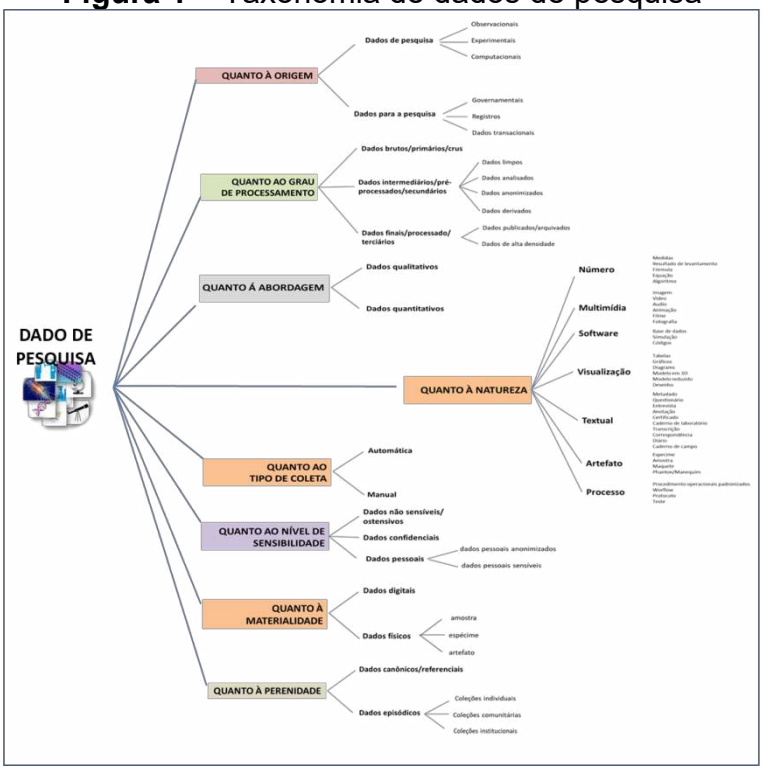
QUANTO À FORMA DE COLETA – Os dados de pesquisa podem ser coletados em qualquer localização – na superfície da Terra, no fundo dos oceanos, na atmosfera ou no espaço interestelar. Isso se dá por instrumentos automatizados, como sensores remotos ou procedimentos manuais, como coleta de artefatos arqueológicos ou entrevistas.

As categorias delineadas podem servir de base para a estruturação de uma taxonomia de dados de pesquisa. A figura 1 exibe a estrutura proposta por Sales e Sayão (2019), como uma síntese do que foi discutido no presente estudo.

[3] Disponível em: https://www.ncbi.nlm.nih.gov/genbank/.
[4] Disponível em: http://www.rcsb.org/.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

Figura 1 – Taxonomia de dados de pesquisa



Fonte: SAYÃO ; SALES (2019).

5 À GUISA DE CONCLUSÃO

A relação entre dado e informação desencadeia uma infinidade de debates que conectam aspectos tecno-sociais com abordagens epistemológicas, que enriquecem os estudos e debates em Ciência da Informação. Porém, para além dessa discussão está a compreensão prática de dados de pesquisa como elemento central dos sistemas de curadoria e gestão que, por sua vez, já fazem parte imprescindível das e-infraestruturas de pesquisa em ambientes científicos mais avançadas.

As propriedades, origem, requisitos e restrições dos dados de pesquisa definirão o conjunto de serviços e aplicações que os sistemas de gestão de dados poderão oferecer às comunidades científicas e à sociedade como um todo. Isto acontece porque dados de pesquisa não podem ser efetivamente curados, preservados, compartilhados e reusados sem uma forte contextualização de muitas faces, especialmente nas dimensões disciplinares e comunitárias; além do mais, dado de pesquisa é um conceito em evolução no mundo da ciência e também em relação às suas conexões com a vida cotidiana, negócios e governo. Conforme enfatiza Borgman (2015), mesmo as instituições que coletam e fazem curadoria de grandes volumes de dados podem não estabelecer

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 32-51, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.11875

uma definição precisa do que eles aceitam ou não como dados de pesquisa; mesmo para o pesquisador, se torna um desafio definir o que é dado de pesquisa de uma forma transversal aos diversos domínios disciplinares que ele atua. De fato, dado de pesquisa permanece um conceito ambíguo, exigindo que os arquivos, repositórios, centros de dados, se adaptem às novas formas de dados na medida em que eles aparecem.

Como contribuição ao entendimento do que devemos curar numa plataforma de dados de pesquisa, o estudo procurou estabelecer uma sistematização possível – manifestada por uma taxonomia – que possa apoiar a compreensão da natureza dos dados e seus fluxos de origem, processamento e uso, com vista à construção de políticas, estratégias e infraestruturas voltadas para a curadoria.

REFERÊNCIAS

ANDERSON, Chris. The end of theory: the data deluge makes the scientific method obsolete. *Wired*, 2008. Disponível em: https://www.wired.com/2008/06/pb-theory/. Acesso em: 25 mar. 2019.

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

BORGMAN, Christine L. Research data: who will share what, with whom, when, and why? In: CHINA-NORTH AMERICAN LIBRARY CONFERENCE, 5., Beijing, 2010. *Proceedings...* Beijing: CALA, 2010. Disponível em: http://works.bepress.com/borgman/238/. Acesso em: 15 maio 2016.

BORGMAN, Christine L. *Scholarship in the digital age*: Information, infraestructure and the internet. London : The MIT Press, 2007.

BORGMAN, Christine L. The conundrum of sharing research data. *Journal of the Association for Information Science and Technology*, v. 63, n. 6, p. 1059-1078, June 2012.

BUCKLAND, Michael K. Information as thing. *Journal of the American Society for Information Science*, v. 42, n. 5, p. 351-360, 1991.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM. *Reference model for an open archival information system (OAIS)*. Washington: Magenta Book, 2012.

GOLD, Anna. Infraestrutura cibernética, dados e bibliotecas, parte 1: Um manual de infraestrutura cibernética para bibliotecários. *D-Lib Magazine,* v. 13, n. 9/10, 2007. Disponível em: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html. Acesso em: 18 ago. 2020.

GOODMAN, Alyssa *et al*. Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, v. 10, n. 4, 2014. Disponível em https://doi.org/10.1371/journal.pcbi.1003542. Acesso em: 18 ago. 2020.

HARVEY, Ross. *Digital Curation*: a how-to-do-it manual. London : Neal-Schuman Publishers, 2010.

JOHNSTON, Lisa. R. *Curating Research Data*: Practical Strategies for Your Digital Repository. Chicago : Association of College and Research Libraries, 2017. Disponível em: https://conservancy.umn.edu/bitstream/handle/11299/185334/Intro_CuratingResearchData_v1.pdf?sequence=1&isAllowed=y. Acesso em: 18 ago. 2020.

KELLEHER, John D.; TIERNEY, Brendan. *Data Science*. Cambridge, MA : MIT Press, 2018.

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION. Appraisal policy of the National Archives. Oct. 2007. Disponível em: https://www.archives.gov/records-mgmt/scheduling/appraisal. Acesso em: 18 ago. 2020.

NATIONAL RESEARCH COUNCIL. *Importance and use of scientific and technical databases*. Washington, DC: The Academic Press, 1999.

NATIONAL SCIENCE BOARD. *Long-lived digital data collections*: enabling research and education in the 21st Century. Arlington: National Science Foundation, 2005. Disponível em: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf. Acesso em: 18 ago. 2020.

NIELSEN, Hans Jørn; HJØRLAND, Birger. Curating research data: the potential roles of libraries and information professionals. *Journal of Documentation*, v. 70, n. 2, p. 221-240, 2014. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JD-03-2013-0034/full/html. Acesso em: 18 ago. 2020.

OECD. Organisation for Economic Co-operation and Development. *OECD*: principles and guidelines for access to research data from public founding. 2007. Disponível em: https://www.oecd.org/sti/sci-tech/38500813.pdf. Acesso em: 19 ago. 2020.

PAMPEL, Heinz *et al*. Making research data repositories visible: the re3data.org Registry. *PLoSOne*, v. 8, n. 11, 2013. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3817176/. Acesso em: 18 ago. 2020.

RESEARCH DATA ALLIANCE. IG physical samples and collections research data ecosystem. In: RDA PLENARY MEETING, 11. *Identifiers for Physical Samples*… Berlin: 2018. Disponível em: https://rd-alliance.org/ig-physical-samples-and-collections-research-data-ecosystem-rda-11th-plenary-meeting. Acesso em: 14 jun. 2019.

SALES, Luana Farias; SAYÃO, Luis Fernando. Uma proposta de taxonomia para dados de pesquisa. *Revista Conhecimento em Ação*, v. 4, n. 1, jan/jun. 2019c. Disponível em: https://revistas.ufrj.br/index.php/rca/article/view/26337. Acesso em: 18 ago. 2020.

SAYÃO, Luis Fernando; SALES, Luana Farias. O fim da teoria: o confronto entre a pesquisa orientada por dados e a pesquisa orientada por hipóteses. *Liinc em Revista*, v. 15, n. 1, p. 16-26, maio 2019a. Disponível em: http://revista.ibict.br/liinc/article/view/4688/4135. Acesso em: 31 jul. 2020.

SAYÃO, Luis Fernando; SALES, Luana Farias. Periódicos de resultados negativos: revelando uma parte invisível da ciência. In: SHITAKU, Milton; SALES, Luana (org.). *Ciência aberta para editores científicos*. Botucatu, SP: ABEC, 2019b. p. 97-102. DOI: http://dx.doi.org/10.21452/978-85-93910-02-9.cap14.

SWANSON, Juleah; RINEHART, Amanda K. Data in context: Using case studies to generate a common understanding of data in academic libraries. *The Journal of Academic Librarianship*, v. 42, n. 1, p. 97-101, 2016. Disponível em: https://kb.osu.edu/bitstream/handle/1811/82202/1/SwansonJ_RinehartA_JAL_Data_in_Context_Preprint.pdf. Acesso em: 18 ago. 2020

THE ROYAL SOCIETY. *Science as an open enterprise***.** London: The Royal Society Science Policy Centre, 2012. Disponível em: https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/.Acesso em: 18 ago. 2020.