ARTIGO

Projeto febre amarela: ciclo de vida e tipologia de dados[1]

Yellow Fever Research Project: lifecycle and data type

Janicy Aparecida Pereira Rocha*

RESUMO

Dados de pesquisa são insumos importantes para a produção de conhecimento. A gestão e a curadoria adequada dos mesmos são formas de se evitar as constantes perdas de dados relatadas na literatura. Assim, apresenta-se a modelagem dos ciclos de vida de dados de pesquisa e a tipificação desses como forma de orientar a promoção de boas práticas de gestão e curadoria. Partindo desses pressupostos, o ciclo de vida dos dados gerados pelo Projeto Febre Amarela – Fiocruz-Minas é modelado e os dados são tipificados. Adicionalmente, são apontadas fragilidades e possibilidades para a gestão dos dados de pesquisa do referido projeto.

Palavras-chave: Dados de Pesquisa; Ciclo de Vida dos Dados de Pesquisa; Gestão dos Dados de Pesquisa.

ABSTRACT

Research data are important inputs for the production of knowledge. The appropriate management and curation of these are ways to avoid the constant loss of data reported in the literature. Thus, this paper presents the modeling of the life cycles of research data and the classification of these as a way of guiding the promotion of good management practices and curation. Based on these assumptions, the data life cycle generated by the Yellow Fever Project - Fiocruz-Minas is modeled and the data are typified. In addition, weaknesses and possibilities for the management of research data of the Yellow Fever project are pointed out.

Keywords: Research Data; Research Data Life Cycle; Research Data Management.

INTRODUÇÃO

Um dos desafios científicos do século XXI é lidar com a avalanche de dados produzidos a partir da ascensão das Tecnologias Digitais de Informação e Comunicação, de forma que eles sejam explorados integralmente, inclusive por quem não os produziu. Cientistas de vários domínios disciplinares usam computadores e diversos outros dispositivos digitais não apenas para ampliar seus esforços tradicionais, mas também para criar e manter redes de interação e colaboração. Isso oportuniza que dados gerados como insumos primários para uma pesquisa sejam, também, subsídios para estudos futuros, inclusive em diferentes projetos.

[1] Agradecimentos são devidos à Fapemig, pelo fomento à pesquisa da qual se deriva esse artigo.

* Doutora em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG). Professora do Departamento de Processos Técnico-Documentais | Centro de Ciências Humanas e Sociais | Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Endereço: Avenida Pasteur, 458, Urca, CEP 22290-240, Rio de Janeiro, RJ. Telefone: (21) 2542-1717. E-mail: janicy.rocha@unirio.br

Ainda que não sejam elementos novos no processo de produção do conhecimento, os dados de pesquisa, agora gerados e usados de forma massiva e distribuída, assumem protagonismo no denominado quarto paradigma científico – *eScience* ou ciência orientada por dados (HEY; TANSLEY; TOLLE, 2009; BORGMAN, 2010). Essenciais para atividades que visam à geração de conhecimento científico e tecnológico, eles são insumos para análises primárias e também derivadas. Nesse reordenamento do fazer científico, os dados, até então pospostos no processo de produção do conhecimento, assumem centralidade, emergindo como ativos valiosos e reutilizáveis.

Há mais de 15 anos, o Relatório Atkins descrevia as potencialidades que os avanços da tecnologia computacional trariam para o fazer científico. Dentre elas, vislumbrava-se a combinação de dados brutos oriundos de várias fontes aliada ao uso de ferramentas para análise, visualização e simulação de inter-relações complexas; o fazer científico extrapolando os limites disciplinares tradicionais – p. ex.: físicos usando diretamente as observações astronômicas – e a possibilidade de acesso *online* a todo o registro científico publicado (ATKINS, *et al*., 2003).

Atualmente, esse cenário atingiu tamanha complexidade que já pode ser entendido como irreversível. Por um lado, é notório que coleções de dados cruciais para um ou mais domínios do conhecimento podem estar agora *online* e remotamente acessíveis, graças às possibilidades provindas dos avanços tecnológicos. Por outro lado, desses constantes avanços originam-se incontáveis desafios não apenas para compartilhamento e abertura das coleções de dados, mas, principalmente, para que elas sejam preservadas em longo prazo, em estruturas adequadas e que possibilitem, ainda, recuperação facilitada e efetiva.

Após analisar 516 artigos e seus respectivos conjuntos de dados, Vines e outros (2014) identificaram que a disponibilidade dos dados diminuía em 17% ao ano após a publicação dos artigos nos quais foram primariamente utilizados. Os participantes do estudo apontaram a obsolescência tecnológica como a principal causa para a perda de dados. Relataram, ainda, perdas causadas pelo roubo de equipamentos ou pelo arquivamento dos dados em mídias externas guardadas em locais distantes, inacessíveis ou já desconhecidos.

Para os referidos autores, esses resultados reforçam a hipótese de que, em longo prazo, pesquisadores não conseguem, individualmente, preservar seus dados de pesquisa de forma adequada. Adicionalmente, muitos pesquisadores, inclusive no Brasil, não dispõem de conhecimento e infraestrutura adequada para a gestão de dados de pesquisa, preferindo armazená-los em arquivos de papel, computadores pessoais, mídias externas e, em menor quantidade, na nuvem (BAUER *et al.*, 2015; VANZ, *et al.*, 2019).Outros tantos pesquisadores não se sentem confortáveis com as mudanças trazidas por esse cenário e preferem recorrer, ainda, à transferência de dados via *email*, mídias removíveis e serviços *online* de compartilhamento de arquivos (EYNDEN *et al*., 2011).

Essas práticas, já instituídas e propagadas entre diferentes gerações de pesquisadores, tendem a disseminar uma cultura de perda frequente de dados de pesquisa. Isso afeta não apenas a reprodutibilidade e a verificabilidade das pesquisas, mas também impede que os dados sejam reutilizados. Em circunstâncias mais extremas, tais práticas impedem até mesmo que os dados sejam utilizados – em situações nas quais eles são coletados e armazenados inadequadamente para análise posterior, tornando-se inacessíveis. Em circunstâncias menos extremas, impera a dificuldade de recuperação e compartilhamento desses dados, resultando em retrabalho e desperdício de tempo e recursos.

Nesse contexto, a gestão de dados de pesquisa se consolida como um campo de estudos no qual os diferentes desafios demandam esforços multidisciplinares. Muitas das respostas a isso tangenciam abordagens advindas especialmente da Ciência da Computação e da Ciência da Informação, sem, no entanto, prescindirem dos aportes de outros domínios do conhecimento. Soma-se a isso a imprescindibilidade de entendimento do domínio no qual os dados se originam

e/ou são utilizados, posto que diferentes domínios possuem diferentes requisitos e, consequentemente, demandam soluções distintas. Não bastassem tais desafios, é preciso, também, conjugar necessidades específicas com padronizações necessárias para possibilitar a interoperabilidade e o reuso.

Ao apresentarem suas contribuições para o estabelecimento de um modelo de curadoria digital para dados de pesquisa para o Brasil, Sayão e Sales (2012) conjugam elementos relativos a aspectos políticos, infraestrutura organizacional, desenvolvimento de coleções de dados, pesquisa, infraestrutura tecnológica e de padronização, formação de recursos humanos, sustentabilidade econômica, implicações sociais, legais e éticas e disponibilização de serviços. Os autores alertam para a necessidade de diálogos e comprometimento entre essas diferentes instâncias para que os desafios sejam superados e os dados de pesquisa possam ser preservados e gerenciados ao longo do tempo, o que favorecerá acesso e reuso.

A curadoria de dados de pesquisa diz respeito ao conjunto de atividades gerenciais, técnicas e informacionais fortemente padronizadas que pressupõe adicionar valor aos dados, durante todo o seu ciclo de vida, para uso corrente e futuro (SAYÃO; SALES, 2012).Como parte de seus ciclos de vida, os dados de pesquisa são coletados, analisados, publicados e, eventualmente, reutilizados e/ou descartados. Entender como esse ciclo se constitui em domínios disciplinares específicos é crucial para o desenvolvimento de orientações e serviços de curadoria capazes de facilitar a preservação, a recuperação e o compartilhamento desses dados pelos pesquisadores.

Nesse artigo são apresentados resultados derivados de uma pesquisa[2] cujo objeto empírico localiza-se nas Ciências da Saúde: um projeto de pesquisa que investiga os mecanismos da resposta imune envolvidos em processos de imunoprofilaxia para febre amarela, doravante Projeto Febre Amarela, desenvolvido no Instituto René Rachou (IRR), unidade regional da Fundação Oswaldo Cruz (Fiocruz), localizada na cidade de Belo Horizonte, Minas Gerais.

Na Fiocruz, a discussão acerca da gestão de dados de pesquisa está atrelada à discussão sobre a abertura dos mesmos, evoluindo em consonância com o movimento da Ciência Aberta. Apesar disso, o "Termo de Referência - Gestão e Abertura de Dados para Pesquisa na Fiocruz" pondera sobre a abertura de dados e a necessidade de sigilo em determinadas situações que visam a sustentabilidade da pesquisa e da instituição (FIOCRUZ, 2018). Assim, embora advogue em favor da abertura dos dados de pesquisa, o referido termo aponta algumas razões como causadoras de restrições ao compartilhamento: questões relativas à segurança ou privacidade de entidades e/ou indivíduos envolvidos na pesquisa; propriedade intelectual; condicionalismos legais diversos, entre outras.

Não obstante o objetivo mais amplo de abertura dos dados de pesquisa, a Fiocruz tem se articulado em torno da temática da gestão de dados de pesquisa, "visando a garantia da integridade, segurança e qualidade dos dados" (FIOCRUZ, 2018, p. 6) e sua preservação em longo prazo. No caso específico aqui tratado, o do Projeto Febre Amarela, a recuperação, o uso e o compartilhamento dos dados gerados entre os próprios integrantes do projeto tem exigido vasto esforço. Os diferentes subprojetos a ele vinculados geram semanalmente, conjuntos de dados que precisam ser reunidos para processamento e análise posteriores. Esses conjuntos de dados assumem abrangência e tamanho consideráveis ao longo do tempo. Além disso, subconjuntos de dados já processados, frequentemente, são armazenados de forma fragmentada em mídias pessoais, resultando em complexo trabalho de recuperação nos momentos de análise e uso.

[2] A pesquisa em questão refere-se à tese de doutorado da autora (ROCHA, 2018), cujo objetivo foi investigar as práticas informacionais relacionadas à produção colaborativa do conhecimento científico e tecnológico em um grupo de pesquisa.

Diagnóstico realizado durante a pesquisa da qual se deriva esse artigo identificou a preocupação de alguns dos colaboradores do Projeto Febre Amarela com a gestão dos dados (ROCHA, 2018). Aliado a isso, nesse diagnóstico também foi identificado o desconhecimento, pela maioria dos colaboradores, das possibilidades e iniciativas atualmente existentes para a gestão de dados de pesquisa, inclusive no âmbito da própria instituição. Não obstante, coordenadores do projeto se mostraram receptivos às possibilidades de gestão para os referidos dados, entendendo-a como necessária.

Frente ao exposto, infere-se o longo caminho que ora se desenha para que a gestão dos dados do Projeto Febre Amarela se torne, de fato, uma realidade. Os resultados aqui apresentados derivam-se de passos iniciais, porém profícuos, nesse sentido. Assim, objetiva-se no presente artigo: (i) modelar o ciclo de vida dos dados gerados no Projeto Febre Amarela; (ii) descrever e analisar as etapas componentes do ciclo de vida modelado, conforme suas particularidades; e (iii) caracterizar os dados de pesquisa produzidos e armazenados, classificando-os conforme seus tipos.

DADOS DE PESQUISA:

Uma simples busca pelo termo "dados de pesquisa" retorna várias definições. À medida que tal busca é aprofundada, mais diversidade surge, conforme se pode perceber em algumas definições apresentadas a seguir:

> [...] qualquer informação que possa ser armazenada em formato digital, incluindo textos, números, imagens, vídeos ou filmes, áudio, *software*, algoritmos, equações, animações, modelos, simulações, etc. Tais dados podem ser gerados por vários meios, incluindo observação, computação ou experimento. (NSF, 2005, p. 13, tradução nossa).
>
> [...] registros factuais (numéricos, textuais, imagens e sons) utilizados como fontes primárias para a pesquisa científica, comumente aceitos na comunidade científica como necessários para validar os resultados da pesquisa. (OCDE, 2007, p.13, tradução nossa).
>
> Dados de pesquisa são todas as informações que foram coletadas, observadas, geradas ou criadas para validar os resultados da pesquisa original. Embora geralmente digitais, os dados de pesquisa também incluem formatos não digitais, como cadernos de laboratório e agendas. (UNIVERSITY OF LEEDS, 2019, *online*, tradução nossa).

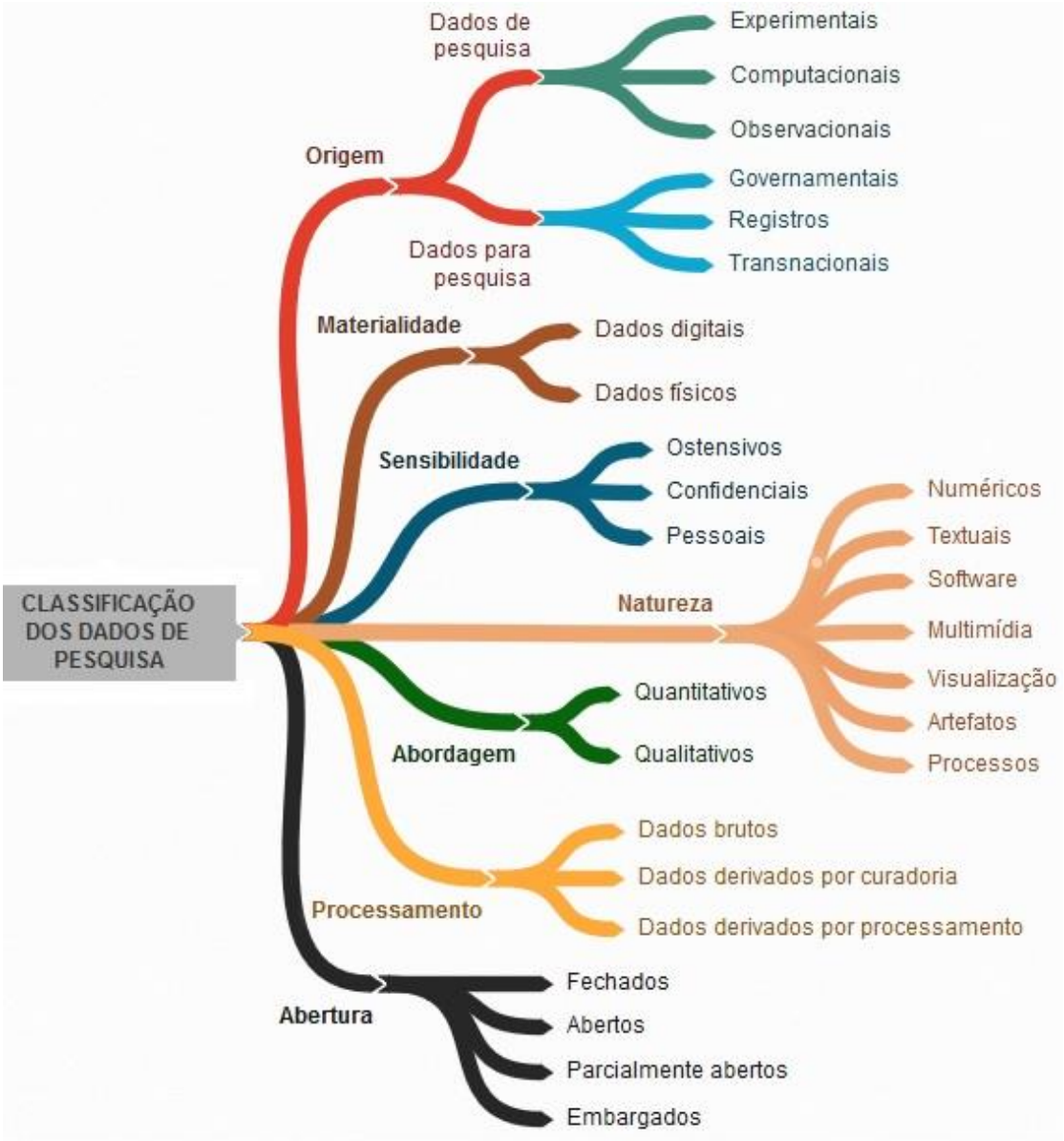
Essa desarmonização conceitual é semelhante àquela existente entre dado, informação e conhecimento. Borgman (2011) ressalta isso, além de apontar que definições explícitas do termo ou menções às variadas formas que eles podem assumir ainda não são uma constante. Para a autora, o entendimento do que seriam os dados de pesquisa é diretamente influenciado por áreas, propósitos e métodos da coleta. Partindo desses alertas, Sales e Sayão (2019, p. 35) concluem que "[...] o que dificulta atribuir uma definição consensual ao dado de pesquisa é o fato idiossincrático que ele pode ser muitas coisas diferentes para pessoas e circunstâncias diferentes [...]", portanto, a definição é dependente de interpretação. Na sequência, os referidos autores propõem a seguinte definição:

> Dado de pesquisa é todo e qualquer tipo de registro coletado, observado, gerado ou usado pela pesquisa científica, tratado e aceito como necessário para validar os resultados da pesquisa pela comunidade científica. (SALES; SAYÃO, 2019, p. 36)

O exposto evidencia que os dados de pesquisa são diversos, de fato. São muitas as variações nas tentativas de defini-los, todavia alguns pontos em comum se destacam – p. ex.: origem, natureza, materialidade, níveis de processamento e sensibilidade, abordagem da pesquisa, entre outros. A

partir disso, algumas categorias genéricas podem ser identificadas e há, na literatura, iniciativas para sintetizá-las. Dentre essas iniciativas, existem classificações mais genéricas e outras mais individuais, que consideram particularidades de domínios específicos. Na Figura 1, são sintetizadas algumas das classificações possíveis, conforme levantamento bibliográfico que considerou aquelas apontadas por Lyon (2007); OCDE (2007); NSF (2007); Borgman (2010); Oliver e Harvey (2017) e Sales e Sayão (2019).

Figura 1 – Classificações de dados de pesquisa



Fonte: Elaborado pela autora, com base na literatura consultada.

Sales e Sayão (2019) defendem que caracterizar os dados de pesquisa é o primeiro passo para a curadoria, já que isso permite identificar metadados peculiares ao tratamento de cada tipo de dado. Para os autores, o tratamento dos dados deve preceder a definição de infraestruturas tecnológicas, bem como o compartilhamento e a abertura dos dados. Identificar a origem dos dados é fundamental, já que produzi-los ou coletá-los novamente pode não ser possível e, portanto, devem ser preservados para sempre. Já a forma pela qual os dados se materializam determina o tratamento que precisam receber. O nível de sensibilidade está diretamente

relacionado às possibilidades de abertura. A natureza demonstra a heterogeneidade dos dados possíveis de serem gerados, em quaisquer das abordagens. Conhecer o nível de processamento é importante tanto para o versionamento quanto para estabelecer ligações entre diferentes versões. Por fim, o grau de abertura determina possibilidades de reuso conforme a sensibilidade dos dados.

Ciclo de vida de dados de pesquisa

Modelos de ciclo de vida de dados de pesquisa têm sido propostos e adotados por diferentes organizações – como Data Observation Network for Earth (DataONE)[3], Digital Curation Centre (DCC)[4], Data Documentation Initiative (DDI)[5] e outras – com o intuito de modelar e descrever a sequência de estágios pelas quais os dados de pesquisa podem passar ao longo de sua existência. Adicionalmente, tais modelos são utilizados para orientar a promoção de boas práticas de gestão, organização e preservação de dados de pesquisa durante cada uma das fases de seus ciclos de vida.

Sobre a necessidade e a utilidade de se modelar tais ciclos de vida, Carlson (2014) argumenta que os dados, ao longo do tempo, provavelmente sofrerão múltiplas transformações tanto no formato, quanto na aplicação e uso e, talvez, até mesmo na finalidade. Assim, ciclos de vida são estruturas úteis para identificar e nomear esses diversos estágios e suas possíveis transformações, articulando-os de forma a contextualizar e comunicar que tipos de serviços de dados podem ser fornecidos, como, para quem e quando.

Atualmente, existe ampla gama de modelos de ciclo de vida de dados de pesquisa, sendo que, muitas vezes, o que os diferencia ou aproxima é o foco ou perspectiva adotados, e classificações relativas a isso começam a surgir. Exemplos são a classificação conforme a estrutura visual – linear, não-linear, circular, entre outras (WISSIK; ĎURČO, 2015) e a classificação conforme o contexto – modelos baseados em indivíduos, relativos a um projeto em específico; modelos baseados em organizações, como universidades, bibliotecas e outras e, finalmente, os modelos baseados em comunidades, direcionados a determinada disciplina ou comunidade acadêmica (CARLSON, 2014).

A despeito da diversidade de atividades, estágios ou etapas do ciclo de vida dos dados de pesquisa, Sayão e Sales (2015) ressaltam que, dependendo do projeto de pesquisa, apenas parte de um modelo de ciclo de vida pode ser utilizada, e etapas que não se aplicam podem ser desconsideradas. Para os referidos autores, alguns modelos tornaram-se referências para pesquisadores, bibliotecários e gestores de dados, sendo adaptados conforme necessidade de cada projeto ou domínio de conhecimento.

Gupta e Müller-Birn (2018) realizaram um trabalho por meio do qual identificaram os principais modelos de ciclo de vida de dados de pesquisa considerados de referência e os compararam quanto aos estágios que abrangem. A síntese dos modelos comparados pelo referido trabalho é apresentada no Quadro 1. Os autores concluem que, nos modelos analisados, se sobressaem os seguintes estágios, contemplados pela maioria: concepção/*design*, coleta, processamento, compartilhamento/distribuição e análise.

[3] https://www.dataone.org/

[4] http://www.dcc.ac.uk/

[5] https://www.ddialliance.org/

Quadro 1– Síntese dos principais modelos de ciclo de vida de dados de pesquisa

MODELO	ORIGEM	ESCOPO PRETENDIDO	ESTÁGIOS	REFERÊNCIA
ICPSR	Inter-university Consortium for Political and Social Research (ICPSR): consórcio mundial de instituições acadêmicas e organizações de pesquisa.	Pesquisa Política e Social	Desenvolvimento de Proposta e Plano de Gestão de Dados; Início do Projeto; Coleta de Dados e Criação de Arquivos; Análise de Dados; Preparação de Dados para Compartilhamento; Depósito de dados.	ICPSR (2012)
DDI	Data Documentation Initiative (DDI): padrão internacional para descrever pesquisas, questionários, arquivos de dados estatísticos e informações de estudos em ciências sociais.	Dados Estatísticos e de Ciências Sociais	Concepção do estudo; Coleta de dados; Processamento de Dados; Compartilhamento de Dados; Descoberta de Dados; Análise de Dados; Reuso; Arquivamento de dados.	DDI (2011)
USGS	Community for Data Integration (CDI): comunidade de prática relativa a dados de pesquisa vinculada ao United States Geological Survey (USGS).	Dados Científicos e Gestão da Informação na Pesquisa em Ciências da Terra	Planejar; Adquirir; Processar; Analisar; Preservar; Publicar/Compartilhar.	USGS (2013)
HUMPHREY'S	Chuck Humphrey, diretor da Portage Network, Associação Canadense de Bibliotecas de Pesquisa.	*eScience* e PesquisaColaborativa	Concepção e *Design* da Pesquisa, Coleta de Dados, Processamento de Dados, Acesso e Disseminação de Dados, Análise, Ciclo KT, Resultados de Pesquisa, Descoberta de Dados, Reutilização de Dados, Dados.	HUMPHREY (2006)
DCC	Digital Curation Conference (DCC): centro voltado	Preservação e Curadoria de Dados	**Ações sequenciais:** Conceituar, Criar ou Receber; Avaliar e	HIGGINS (2008)

	para os desafios da preservação digital e da curadoria digital nas instituições de ensino superior do Reino Unido.		Selecionar; Inserir; Preservar; Armazenar; Acessar; Usar; Reutilizar; Transformar. **Ações ocasionais:** Descartar; Reavaliar; Migrar.	
IWGDD	Interagency Working Group on Digital Data (IWGDD): grupo de Agências Federais nos EUA para promoção de estruturas interoperáveis abertas para dados digitais de Pesquisa e Desenvolvimento (P&D).	Preservação e Acesso de Dados para P&D	Planejar; Criar; Manter; Distribuir.	IWGDD (2009)
DATAONE	Data Observation Network for Earth (DataONE): plataforma para a ciência ambiental e ecológica, conduzida pela comunidade mundial.	Dados de Observação da Terra para Pesquisa Ambiental e Ecológica	Coletar; Assegurar; Descrever; Depositar; Preservar; Descobrir; Integrar; Analisar.	DATAONE (2013)

Fonte: Elaborado pela autora, com base em Gupta e Müller-Birn (2018)

Gestão de Dados de Pesquisa

Em 2009 a Revista Nature já destacava, no editorial *Data's Shame ful Neglect*, a gestão de dados de pesquisa como um dos alicerces da produção científica contemporânea: "A pesquisa não pode florescer se os dados não forem preservados e estiverem acessíveis." (NATURE, 2009, p. 145, tradução nossa). Por gestão de dados de pesquisa entendem-se as várias atividades associadas às diversas fases do ciclo de vida, voltadas para planejamento, obtenção, armazenamento, segurança, preservação, recuperação, compartilhamento e reutilização dos dados, além de considerar as capacidades técnicas, considerações éticas, questões jurídicas e estruturas de governança relacionadas (WHYTE; TEDDS, 2011; COX; PINFIELD, 2013).

Frente à importância e à longevidade dos dados de pesquisa, gerenciá-los é fundamental e, para tanto, é necessário planejamento. O Plano de Gestão de Dados (PGD) consiste em um "[...] documento formal que estabelece um compromisso de como esses dados serão tratados durante todo o desenvolvimento do projeto, e também após a sua conclusão." (SAYÃO; SALES, 2015, p. 15). Ele torna-se cada vez mais necessário, sendo exigido por algumas agências públicas e privadas de fomento à pesquisa científica na América do Norte, Europa, Austrália e, mais recentemente, do Brasil. Para os autores supracitados, o PGD deve ser um documento dinâmico que, embora baseado em elementos comuns, possa ser adaptado ao longo do projeto.

Um desses elementos é a descrição minuciosa de tópicos como o contexto no qual os dados foram criados, o contexto tecnológico de geração dos arquivos de dados, os instrumentos usados na coleta, parâmetros de qualidade e outros. Para isso, o uso de um padrão de metadados bem definido é determinante, facilitando o rastreamento, o uso e o reuso dos dados.

Políticas de acesso, compartilhamento e reuso também devem ser consideradas no PGD, com atenção especial para questões éticas, segurança e proteção dos dados e *copyright*. A gestão do arquivamento de longo prazo também deve ser incluída no PGD de forma que os conteúdos digitais sejam preservados "[...] mantendo as suas características de autenticidade, integridade e proveniência, de forma que eles estejam sempre disponíveis e prontos para serem usados." (SAYÃO; SALES, 2015, p. 24). Para a preservação de conteúdos digitais, repositórios e centros de dados são opções mais adequadas que mídias sujeitas à obsolescência tecnológica.

PERCURSO METODOLÓGICO

A presente pesquisa se caracteriza como essencialmente qualitativa; sendo exploratória e descritiva, quanto aos seus objetivos. Observações e entrevistas contextuais foram adotadas como técnicas de coleta de dados. Originária da Investigação Contextual (HOLTZBLATT, JONES; 1993), a entrevista contextual consiste em conversas para esclarecimentos realizadas em conjunto com a observação. A escolha por essa modalidade de entrevista deveu-se ao fato de as observações acontecerem durante a execução de atividades rotineiras dos colaboradores do projeto, em situações que não permitiam longas pausas para conversas. Parte-se do pressuposto de que os participantes possuem conhecimento superior sobre o ambiente empírico e os pesquisadores possuem elevado interesse em compreendê-lo. Complementarmente, foram consultadas algumas coleções de dados do projeto, de diferentes naturezas e níveis de processamento.

Para que os colaboradores do Projeto Febre Amarela fossem informados da presença da autora dessa pesquisa em suas dependências, a mesma se apresentou a eles durante uma reunião do grupo de pesquisa no qual o referido projeto é executado. Posteriormente, por intermédio da coordenadora do projeto, foi enviado aos participantes em potencial um *email* em formato de carta convite, com informações adicionais sobre a pesquisa. Colaboradores que não responderam ao e-mail foram convidados pessoalmente pela autora da pesquisa. Aqueles que alegavam algum fator dificultador para o agendamento foram contatados por mais três vezes e, não sendo possível o agendamento, ou tendo se recusado explicitamente a participar da pesquisa, foram removidos da lista de participantes.

Aquele que manifestaram consentimento em participar da pesquisa, assinaram o Termo de Consentimento Livre e Esclarecido (TCLE), elaborado quando da submissão da proposta de pesquisa ao Comitê de Ética da Universidade Federal de Minas Gerais. Assim, foram realizadas observações sistemáticas das atividades de 16 colaboradores do projeto em questão desde a realização de experimentos em laboratório até a redação dos diferentes relatos de pesquisa, além das entrevistas contextuais.

DADOS DE PESQUISA NO PROJETO FEBRE AMARELA: RESULTADOS PARCIAIS

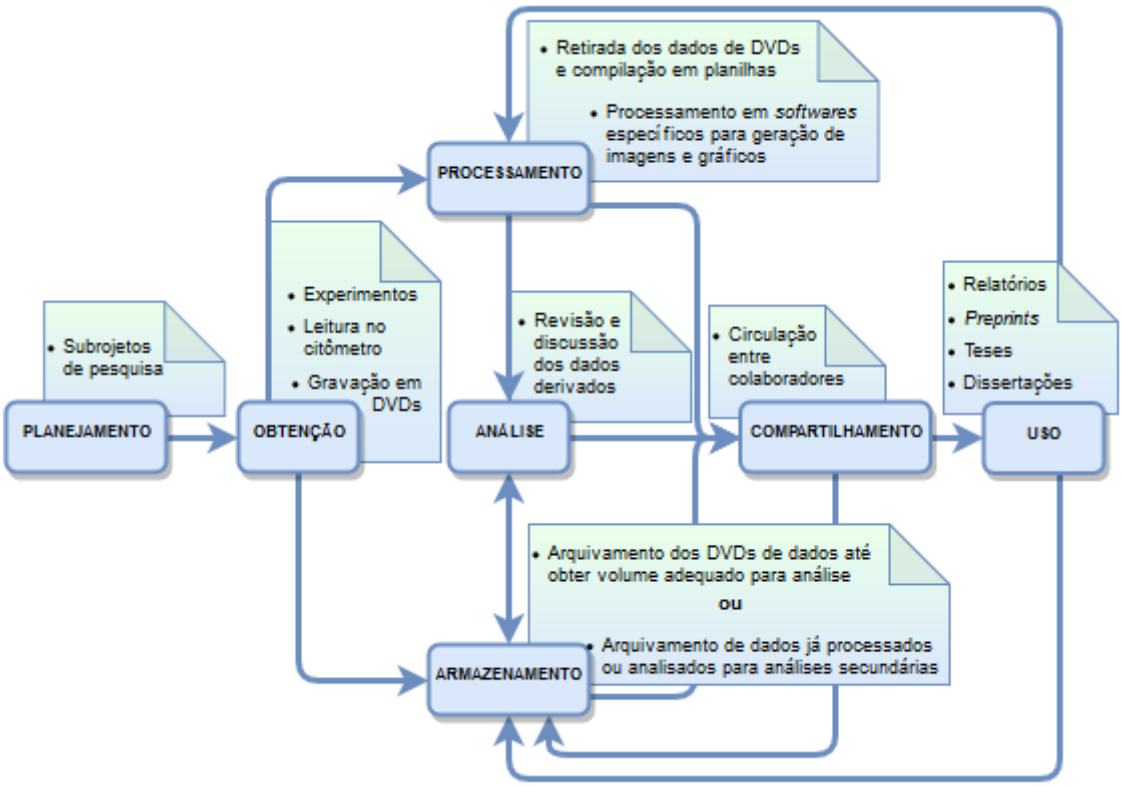
Identificadas as particularidades do processo de produção do conhecimento no Projeto Febre Amarela, procedeu-se à modelagem do ciclo de vida de seus dados de pesquisa. Aliado a isso, foi possível a identificação dos tipos de dados gerados e suas

formas de transformação ao longo de cada etapa do ciclo de vida, o que permitiu classificá-los conforme a tipologia apresentada na Figura 1. Os resultados são apresentados nas seções 4.1 e 4.2.

Ciclo de vida

A compreensão do ciclo de vida dos dados gerados no âmbito do Projeto Febre Amarela inicia-se com o mapeamento e a descrição de suas diversas etapas (Figura 2), a partir da observação das atividades executadas em cada uma delas e tendo como referências as fases dos principais modelos de ciclo de vida de dados de pesquisa, apresentadas no Quadro 1. Concomitantemente à descrição de cada etapa, é feito um diagnóstico preliminar de fragilidades relativas à gestão dos dados gerados no contexto em questão.

Figura 2 – Ciclo de vida dos dados do Projeto Febre Amarela



Fonte: Elaborado pela autora, com base nos dados da pesquisa.

Planejamento da pesquisa

Nessa etapa são preparadas as propostas das diferentes pesquisas derivadas do projeto. Mediante demandas recebidas pelos coordenadores, ou identificadas por quaisquer colaboradores, são realizadas reuniões que resultam em um esboço da pesquisa contendo, geralmente, perguntas, hipóteses, justificativas e métodos. Então, questões práticas de operacionalização da pesquisa são discutidas – por exemplo, tamanho da amostra; perfil e forma de seleção dos voluntários; protocolos a serem usados; tarefas de cada colaborador do projeto; entre outras. Definidas essas questões, elabora-se um projeto de pesquisa, que é submetido ao Comitê de Ética e, se aprovado, inicia-se a etapa de obtenção de recursos financeiros para execução. O PGD não faz parte desse estágio e, questionados, alguns participantes alegaram

desconhecê-lo, bem como sua exigência por agências de fomento. Da mesma forma, não há nenhuma formalização relacionada a políticas de acesso, preservação, segurança, compartilhamento e reuso dos dados.

Obtenção dos dados

Armazenadas em tubos de ensaio, as amostras biológicas são recebidas no laboratório. A partir da interação entre colaboradores e artefatos diversos – quando reagentes são adicionados às amostras e essas são centrifugadas, colocadas em cultura, lavadas e caracterizadas – as representações dos dados sobre a resposta imune de voluntários vão sendo transformadas. De sangue humano em um tubo de ensaio, ela se torna um conjunto de células, caracterizadas conforme o interesse do subprojeto. O equipamento BD LSRFORTESSA é o citômetro de fluxo[6] utilizado para leitura das células caracterizadas no laboratório, o que geralmente ocorre na semana seguinte ao recebimento das amostras.

No início dessa etapa, os dados sobre a resposta imune dos voluntários são representados pela amostras biológicas, em tubos de ensaio. Após os experimentos em laboratórios, os dados são representados pelas células caracterizadas. Após leitura no citômetro, tais células são convertidas em dados brutos (*raw data*) – também chamados de dados primários, assumindo um formato digital e sendo gravados em DVDs por meio de *software* proprietário específico. Os DVDs são identificados pela data do experimento e o título do subprojeto. Nesse momento, os dados, experimentais, possuem natureza numérica, textual e imagética. Informações relativas aos experimentos também são anotadas no Livro de Registro, versão institucional dos também chamados Cadernos de Laboratório.

Processamento dos dados

Sucintamente, esse processo ocorre da seguinte forma: os responsáveis pelo processamento copiam os dados dos DVDs para um HD externo que é conectado ao computador (já que esse não possui *drive* de DVD) e, então, utilizam o *software* Flow Jo[7] para processá-los. Já previamente processados, tais dados são compilados em planilhas eletrônicas e, então, submetidos ao GraphPad Prism[8], *software* para análise estatística e apresentação gráfica de dados científicos, gerando figuras e gráficos. Nenhum metadado é atribuído a esses dados processados.

Armazenamento dos dados

O armazenamento dos dados acontece de duas formas: (i) DVDs com dados brutos são armazenados em uma caixa até a obtenção de volume de dados suficiente para processamento, ou até ter colaborador disponível para realizar tal tarefa; (ii) dados processados são salvos em HD externo ou enviados por email, ficando assim até o

[6] Equipamento por meio do qual é possível avaliar, em partículas microscópicas suspensas em meio líquido, um conjunto de parâmetros previamente definidos.

[7] https://www.flowjo.com/

[8] https://www.graphpad.com/scientific-software/prism/

momento da análise. Uma vez analisados e utilizados em publicações, esses dados permanecem arquivados – e a eles são acrescidos outros dados – para utilização posterior. Nenhuma política oficial de *backup* foi identificada, apenas a duplicação de dados em várias mídias e emails. Também não existem políticas de versionamento ou descarte.

Compartilhamento dos dados

Por compartilhamento entende-se a circulação dos dados processados entre os colaboradores, para subsidiar discussões acerca dos resultados a serem publicados. Os dados são compartilhados via email ou mídias externas, sem controle sobre quais perfis de colaboradores podem ter acesso a diferentes coleções de dados.

Análise dos dados

A análise é entendida como a exploração do compêndio de dados processados, o que acontece em reuniões presenciais, para subsidiar a elaboração de relatórios, artigos, teses e dissertações, gerando resultados relativos a objetivos preestabelecidos e/ou respostas para as questões dos subprojetos.

Uso dos dados

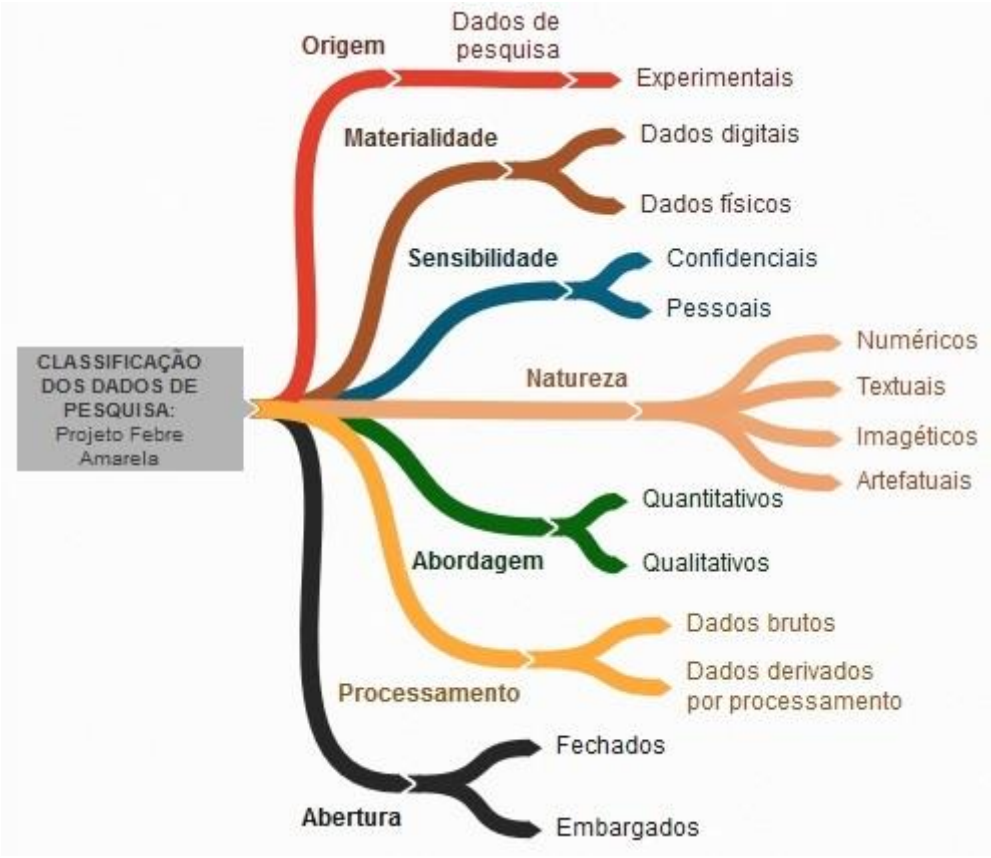
Por uso entende-se a seleção e utilização dos dados previamente discutidos em relatórios, artigos, teses e dissertações. Uma vez escritas, as diferentes versões desses textos circulam entre os colaboradores por email ou em formato impresso.

Tipologia

A classificação dos dados do Projeto Febre Amarela, feita conforme tipologias propostas por Lyon (2007); OCDE (2007); NSF (2007); Borgman (2010); Oliver e Harvey (2017) e Sales e Sayão (2019) é apresentada na Figura 3. No que concerne à origem, os dados são experimentais, provenientes de situações controladas em bancadas de laboratórios. Todavia, é preciso se atentar que, em alguns dos subprojetos, eles não podem ser facilmente reproduzidos uma vez que são coletados de voluntários com quadros clínicos específicos, que podem levar anos para se repetir ou até mesmo não se repetir da mesma forma.

Quanto à materialidade, a maioria dos dados é digital, mas também existem dados físicos –p. ex.: amostras e Livro de Registro. Já em relação ao nível de sensibilidade, a princípio, são entendidos como dados pessoais sensíveis e, portanto, confidenciais, na medida em que contêm dados clínicos de voluntários e não passaram por tratamentos de anonimidade criteriosos. Quanto à natureza, são dados numéricos, textuais, imagéticos e artefatuais (amostras biológicas, células caracterizadas, Livro de Registro, DVDs).

Figura 3 – Classificação dos dados do Projeto Febre Amarela



Fonte: Elaborado pela autora, com base nos dados da pesquisa.

Quanto à abordagem, são dados quantitativos e qualitativos. Em termos de processamento, existem dados brutos, mas também aqueles que já foram previamente processados. Finalmente, em relação ao grau de abertura, atualmente, são considerados dados fechados e/ou embargados, já que não passaram por nenhuma atividade de curadoria e gestão e, além disso, ainda não aconteceram discussões profícuas sobre possibilidades de abertura desses dados.

CONSIDERAÇÕES FINAIS

Dados de pesquisa são fundamentais na medida em que consistem em importante insumo para atividades que visam à geração de conhecimento científico e tecnológico. Atualmente, de subprodutos da pesquisa, os dados passam a ativos valiosos não apenas para aquele projeto do qual se originaram. Nesse cenário, a gestão dos dados de pesquisa torna-se cada vez mais necessária, porém não menos desafiadora, apesar das potencialidades trazidas pelas tecnologias. São notórios os benefícios do planejamento antecipado de como os dados de pesquisa serão coletados, documentados, gerenciados e preservados. Apesar disso, ainda prevalece o desconhecimento sobre como fazer a gestão e a curadoria adequadas, especialmente por parte dos pesquisadores envolvidos nos projetos nos quais esses dados são gerados.

No domínio objeto dessa pesquisa, falta infraestrutura tecnológica para gestão de dados e, sobretudo, conhecimento acerca da gestão de dados por parte dos colaboradores do projeto. Apesar disso, foi identificada importante receptividade às possibilidades de ações para gestão e curadoria efetiva dos dados produzidos. São

dados cuja importância no cenário de pesquisas sobre a febre amarela é incontestável: os estudos desenvolvidos pelo projeto geram conhecimento científico orientado à resolução de problemas de saúde pública relevantes e atuais.

Ainda que o apresentado nesse artigo seja resultado de passos iniciais, a prioridade de ações futuras tem sido estabelecida. Urge que os conjuntos de dados sejam documentados, ou seja, que a eles sejam adicionadas metadados que descrevam o contexto de sua origem. Documentar o que os dados significam, qual o conteúdo e a estrutura deles, além de processamentos que possam ter ocorrido é relevante, posto que dados produzidos no âmbito de determinado projeto podem ser utilizados em uma agenda de pesquisa completamente diferente, por outro grupo de pesquisadores ou ainda, pelo mesmo projeto, em outra época e por outra equipe.

Da mesma forma, urge que ações para armazenamento seguro sejam iniciadas, assim como ações para desenvolver, nos colaboradores, competências para a gestão dos dados por eles gerados e mantidos. A abertura desses dados ainda não é objeto de discussões em estágios avançados, embora exista receptividade para isso. Não obstante, a preservação dos mesmos é ponto central e precisa avançar de forma consistente e ágil.

Artigo recebido em 05/07/2019 e aprovado em 04/11/2019.

REFERÊNCIAS

ATKINS, D. E. *et al*. *Revolutionizing science and engineering through cyberinfrastructure:* report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure. [*S.l.: s.n.*], 2003.

BAUER, B. *et al*. *Researchers and their data:* results of an Austria survey: report 2015. 2015. Disponível em: https://phaidra.univie.ac.at/o:409318. Acesso em: 30 maio 2019.

BORGMAN, C. L. Research data: who will share what, with whom, when a why. *RatSWD Working Paper,* v. 161, n. 10, 2010. Disponível em: http://sydney.edu.au/research/data_policy/resources/ands_borgman_2010_research_data.pdf. Acesso em: 28 maio 2019.

CARLSON, J. The use of data lifecycle models in developing and supporting data services. *In:* RAY, J. M. *Research data management:* practical strategies for information professionals. Purdue University Press, 2014. p. 63-86.

COX, A. M.; PINFIELD, S. Research data management and libraries: current activities and future priorities. *Journal of Librarianship and Information Science,* v. 46, n. 4, p. 299-316, 2013.

DATAONE. *Data management guide for public participation in scientific research:* DataONE public participation in Scientific Research Working Group, Albuquerque. [S.l.: s.n.], 2013.

DDI. **The Data Documentation Initiative:** an introduction for national statistical institutes. 2011. p. 1-10. Disponível em: http://odaf.org/papers/DDI_Intro_forNSIs.pdf. Acesso em: 28 jun. 2019.

FIOCRUZ. Grupo de Trabalho em Ciência Aberta. *Termo de referência:* gestão e abertura de dados para pesquisa na Fiocruz. Rio de Janeiro: FIOCRUZ/Presidência,

2018. 15p. Disponível em: https://www.arca.fiocruz.br/handle/icict/26803. Acesso em: 30 maio 2019.

GUPTA, S.; MÜLLER-BIRN, C. A study of e-Research and its relation with research data lifecycle: a literature perspective. *Benchmarking,* v. 25, n. 6, p. 1656-1680, 2018.

HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). *The fourth paradigm:* data-intensive scientific discovery. Microsoft Research, 2009. Disponível em: http://research.microsoft.com/collaboration/fourthparadigm. Acesso em: 12 maio 2019.

HIGGINS, S. The DCC curation lifecycle model**.** *The International Journal of Digital Curation,* v. 3, n. 1, p. 134-140, 2008.

HOLTZBLATT, K.; JONES, S. Contextual inquiry: a participatory technique for system design. *In:* SCHULER, D.; NAMIOKA, A. (ed.). *Participatory design:* principles and practices. Hillsdale, NJ: Lawrence Erlbaum Associates. 1993, p. 177-210.

ICPSR. *Guide to social science data preparation and archiving:* best practice throughout the data lifecycle. Michigan: Inter-university Consortium for Political and Social Research (ICPSR); Institute for Social Research University of Michigan, Ann Arbor, 2012.

IWGDD. *Harnessing the power of digital data for science and society:* report of the Interagency Working Group on Digital Data (IWGDD) to the Committee on Science of the National Science and Technology Council. [*S.l.: s.n.*], 2009.

LYON, L. *Dealing with data:* role, rigths, responsabilities and relationships. Consultancy report, p. 1-65, 2007. Disponível em: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf. Acesso em: 03 jul. 2019.

NATURE. Data's shameful neglect. *Nature,* v. 461, n. 7261, p. 145, 10 Sept. 2009.

NATIONAL SCIENCE BOARD. *Long-lived digital data collections:* enabling research and education in the 21st century. Arlington, VA: National Science Foundation, 2005.

ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. *Principles and guidelines for access to research data from public data.* 2007. Disponível em: http://www.oecd.org/dataoecd/9/61/38500813.pdf. Acesso em: 17 jun. 2019.

OLIVER, G; HARVEY, R. *Digital curation.* Chicago: American Library Association, 2017.

ROCHA, J. A. P. *A produção do conhecimento como cognição distribuída:* práticas informacionais no fazer científico. 2018. 210 f. Tese (Doutorado) - Universidade Federal de Minas Gerais, Escola de Ciência da Informação. Disponível em: http://hdl.handle.net/1843/BUOS-B3UK96. Acesso em: 03 nov. 2019.

SALES, L. F.; SAYÃO, L. F. Uma proposta de taxonomia para dados de pesquisa. *Conhecimento em Ação,* Rio de Janeiro, v. 4, n. 1, 2019. Disponível em: https://revistas.ufrj.br/index.php/rca/article/view/26337/14573. Acesso em: 08 jul. 2019.

SAYÃO, L. F.; SALES, L. F. CURADORIA DIGITAL: um novo patamar para preservação de dados digitais de pesquisa. *Informação e Sociedade:* Estudos, João Pessoa, v. 22, n. 3, p.179-191, set./dez. 2012.

SAYÃO, L. F.; SALES, L. F. *Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores.* Rio de Janeiro: CNEN/IEN, 2015. Disponível em: http://www.cnen.gov.br/images/CIN/PDFs/GUIA_DE_DADOS_DE_PESQUISA.pdf. Acesso em: 22 jun. 2019.

UNIVERSITY OF LEEDS. *Research data management explained.* Disponível em: https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained. Acesso em: 30 jun. 2019.

USGS. *The United States geological survey science data lifecycle model:* open-file Report nº 2013-1265, US Geological Survey (USGS), Reston, VA. 2013.

VAN DEN EYNDEN, V. *et al*. *Managing and sharing data:* a best practice guide for researchers. United Kingdom: Data Archive, University of Essex, 2011.

VANZ, S. A. de S. *et al*. *Acesso aberto a dados de pesquisa no Brasil:* práticas e percepções dos pesquisadores: relatório 2018. Disponível em: http://hdl.handle.net/10183/185195. Acesso em: 30 maio 2019.

VINES, T. H. *et al*. The availability of research data declines rapidly with article age. *CurrentBiology,* v. 24, n. 1, p. 94-97, 2014. Disponível em: https://www.cell.com/action/showPdf?pii=S0960-9822%2813%2901400-0. Acesso em: 03 jun. 2019.

WHYTE, A.; TEDDS, J. *Making the case for research data management:* Digital Curation Centre Briefing Papers. Edinburgh: JISC, 2011. Disponível em: http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm. Acesso em: 2 jul. 2019.

WISSIK, T.; ĎURČO, M. Research data workflows: from research data lifecycle models to institutional solutions. *CLARIN AnnualConference2015,* Poland, v. 123, p. 94-107, 2016. Disponível em: http://www.ep.liu.se/ecp/123/008/ecp15123008.pdf. Acesso em: 03 jun. 2019.