Identificando dados de pesquisa nas humanidades

Identifying research data in the humanities / Identificación de datos de investigación en humanidades

Márcia Cavalcanti
Doutora em Ciência da Informação, pelo convênio do Instituto Brasileiro de Informação em Ciência e Tecnologia e da Universidade Federal do Rio de Janeiro (Ibict/UFRJ). Professora do curso de Administração da Universidade Santa Úrsula (USU), Brasil.
marciacavalcanti@gmail.com

RESUMO

A partir da revisão de literatura, percebeu-se que não existe consenso sobre o termo dados de pesquisa, nem definição que atenda principalmente ao campo das humanidades digitais. Assim, o artigo apresenta uma proposta de padronização dos dados de pesquisa em humanidades, na forma de uma classificação tipológica, visando auxiliar na elaboração do Plano de Gestão de Dados (PGD) e no processo de depósito para preservação em repositórios.

Palavras-chave: dados de pesquisa; gestão de dados; humanidades digitais.

ABSTRACT

From the literature review it was realized that there is no consensus on research data, neither definition especially in the field of digital humanities. So, the article presents a proposal for standardizing for research data in the humanities, in the form of a typological classification to assist in the elaboration of the Data Management Plan (DMP) and in the deposit process for preservation in repositories.

Keywords: research data; data management; digital humanities.

RESUMEN

A partir de la revisión de la literatura, se observó que no existe un consenso sobre los datos de la investigación y tampoco una definición que se ajuste as necesidades especialmente en el campo de las humanidades digitales. Así, el artículo presenta una propuesta de estandarización de los datos de investigación en humanidades en forma de clasificación tipológica con el objetivo de ayudar en la elaboración del Plan de Gestión de Datos (PGD) y en el proceso de depósito para su preservación en repositorios.

Palabras clave: datos de investigación; gestión de datos; humanidades digitales.

Introdução

Fox e Hendler (2011, p. 159) apontam que "as tecnologias tradicionais não foram projetadas para lidar com a escala e a heterogeneidade de dados no mundo moderno". Não obstante, sabe-se que, no domínio da ciência, diversos autores já nomeiam a época em que vivemos como a da pesquisa científica centrada em dados. Ou seja, um período em que se entendem e se definem, como dados de pesquisa (*research data*), aqueles que são gerados pelos pesquisadores no desenvolvimento de suas pesquisas e que podem ser, posteriormente, reutilizados pelos pares, ou por outros além da comunidade acadêmica.

Acompanhando tal tendência, hospedada no domínio da University of Leicester, está uma página intitulada What is Research Data. Nela é possível acessar um documento de cinco páginas que reúne várias definições e explicações úteis sobre dados de pesquisa, mas que também deixa claro que dar uma definição para o termo é um desafio, principalmente pela falta de consenso entre os diferentes campos do conhecimento. Essa diferença é dinâmica e crescente: "As bases de dados de pesquisa estão crescendo em número e volume, e os diferentes campos da ciência passam a ter acesso a um número mais abrangente de fontes de dados" (Sales; Cavalcanti, 2015, p. 89). Com efeito, as fontes de dados (*data sources*) de que falam Sales e Cavalcanti se traduzem nos suportes onde estão os dados, como uma planilha, um vídeo, uma imagem ou um banco de dados.

Nesse ínterim, a área das humanidades também começa a se voltar para a percepção da importância da preservação dos dados produzidos em suas pesquisas, bem como daqueles provenientes de outras áreas, como os dados demográficos ou econômicos, já que os utiliza também. Como garantir que eles sejam preservados para reúso, ou como forma de comprovação da pesquisa realizada? Essa questão se torna, a cada dia, mais imperativa, em face da aceleração das formas de produção da informação e do conhecimento, inclusive em áreas do conhecimento mais tradicionalmente afastadas de temáticas e objetos próprios da computação, por exemplo. As humanidades pedem urgência para esse debate, principalmente porque a aceleração dessas formas de produção da informação também torna obsoletas suas formas de reprodução.

Uma definição bem primária de dado vem da computação, que o entende como uma sequência de símbolos (letras ou números), os quais, segundo Setzer (2015), podem ser descritos, armazenados e manipulados por computadores. Mas, para que esses dados passem a ter algum significado, eles precisam ser contextualizados. Tal contextualização aumenta potencialmente à medida que os campos ora distintos constroem pontes de diálogo e colaboração em suas ações. Tal cenário parece ser cada vez mais identificável em iniciativas como as da ciência aberta,

que, afinal, pressupõe que os dados estejam abertos. "Dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa – sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras" (Poikola; Villum; Dietrich et al., 2017). A ciência aberta é um modelo de prática científica que, ao contrário dos modelos anteriores, nos quais as informações sobre a pesquisa em andamento, ou já realizada, ficavam restritas a um número limitado de pesquisadores, propõe o compartilhamento da informação em rede. Dessa forma, para que os dados sejam considerados abertos, seus metadados devem ser públicos, ou seja, tem de ser possível poder encontrá-los, mesmo que eles tenham restrições de acesso e de uso.

As três leis dos dados abertos, de David Eaves, propõem que: 1) se o dado não pode ser encontrado e indexado na web, ele não existe; 2) se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado; e 3) se algum dispositivo legal não permitir sua replicação, ele não é útil (Eaves, 2009, tradução nossa).

Para que tudo isso seja concretizado, para que o pesquisador elabore um Plano de Gestão de Dados (PGD) no qual ele informe quais serão os procedimentos para o compartilhamento de seus dados, é preciso caracterizar o que são dados de pesquisa. PGD é um documento exigido pelas principais agências de fomento internacionais, e agora por algumas no Brasil, que identificam a necessidade do tratamento dos dados de pesquisa como fundamental para todas as áreas do conhecimento (Sayão; Sales, 2020; Lima, 2020; Santos; Clinio, 2019).

Neste artigo nos voltaremos para uma discussão sobre o que são dados de pesquisa em humanidades, não como uma tentativa de definir e limitar o termo, mas sim propondo uma padronização na forma de uma classificação tipológica, visando auxiliar os pesquisadores na elaboração de seus planos de gestão de dados e aqueles envolvidos no processo de depósito para preservação em repositórios. "Muito raramente, aqueles que promovem o compartilhamento e a curadoria de dados definem 'dados' explicitamente ou reconhecem a diversidade de formas que esses dados podem assumir", como já pontuou Borgman (2010, p. 3, tradução nossa).

Primeiros passos em um novo território: prospecção do que serão dados de pesquisa em humanidades

Para a realização deste trabalho, foi feito um levantamento bibliográfico de artigos e livros disponíveis em sites acadêmicos e de pesquisa, nacionais e internacionais (Fapesp; NSF; University of Leeds, entre outros), e uma busca

com diferentes termos na ferramenta Google, sem fechar um período temporal específico.

A pesquisa bibliográfica buscou seguir, na medida do possível, a sistemática sugerida por Lakatos e Marconi (2010, p. 26), na qual foram observadas as seguintes etapas: 1) escolha do tema: dados de pesquisa em humanidades, devido à dificuldade de consenso e clareza de sua definição; 2) identificação: pesquisa por referências que abordem o tema do artigo, que se deu pela seleção de palavras-chave para a busca – "dados de pesquisa em humanidades", "dados de pesquisa em ciências humanas", "dados de pesquisa", *"research data"*, *"social science data"*; 3) compilação: as referências foram agrupadas por afinidade em relação ao tema do estudo; 4) fichamento: primeiro foram lidos os resumos do material selecionado para verificar se o conteúdo estava adequado ao escopo da pesquisa, para então seguir com a seleção do material a ser lido; 5) análise e interpretação: leitura crítica das referências selecionadas com interpretação subjetiva das informações textuais recuperadas; e 6) redação: preparada de forma descritivo-textual, contendo citações do material selecionado que embasam as questões defendidas no trabalho.

A página da Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp, s.d.) foi o local de início da busca por sites de universidades e agências de fomento que pudessem contribuir com o tema dados de pesquisa em humanidades.

Partiu-se da discussão e de questionamentos que vêm se consolidando há alguns anos, principalmente a partir dos avanços tecnológicos no campo acadêmico, que envolvem os dados que os cientistas geram/coletam durante o desenvolvimento de suas pesquisas, os chamados dados de pesquisa: o que fazer com eles ao longo e ao final da pesquisa? Como preservá-los? Como proceder com sua gestão?

> A gestão de dados científicos cobre todo o chamado "ciclo de vida" dos dados, ou seja, desde a sua coleta até o armazenamento de longo prazo, passando por uma série de processamentos de limpeza, curadoria, anotação, indexação e transformação. Grande parte da pesquisa científica de hoje exige algum tipo de análise e processamento de dados. Com isto, o planejamento da gestão dos dados utilizados e gerados em uma pesquisa passou a fazer parte integral da metodologia científica, sendo, inclusive, considerado como um dos itens necessários de boas práticas de pesquisa. (Medeiros, 2018)

Que os cientistas geram uma grande quantidade de dados em suas pesquisas não há dúvida, mas nas humanidades algumas questões se tornam urgentes. Uma delas se relaciona à necessidade de se elaborarem critérios que possam definir e identificar os dados gerados pelas pesquisas nas áreas que a compõem

para, assim, poder seguir adiante e responder duas perguntas essenciais sobre a gestão desses dados: 1) que dados serão gerados?; 2) como eles serão preservados e disponibilizados para o reúso?

Um Plano de Gestão de Dados (PGD) passou a ser um documento exigido pelas principais agências de fomento de países da Europa e da América do Norte, e envolve a relação de diferentes itens, dentre eles a descrição de mecanismos, formatos e padrões para armazenar os dados gerados pelas pesquisas e, posteriormente, torná-los acessíveis, podendo incluir até mesmo o uso de repositórios.

> Vale a pena notar que as instruções variam bastante, em função da área do conhecimento. No entanto, o conjunto de informações básicas a ser fornecido é sempre o mesmo – quais dados serão produzidos pelo projeto, restrições de compartilhamento, como serão compartilhados e como serão preservados. (Fapesp, s.d.)

A Fapesp foi a primeira agência de fomento no Brasil a se preocupar com a elaboração de um PGD e a exigi-lo nos pedidos de financiamento de pesquisa, como já vinha ocorrendo em outros países. Em sua página on-line, disponibiliza na área "Fomento à pesquisa" o link Gestão de dados,[1] no qual o pesquisador tem acesso aos itens: 1) gestão de dados; 2) conteúdo do Plano de Gestão de Dados (Fapesp); 3) ferramentas on-line para criação de planos; 4) documentos e páginas de interesse (planos de gestão de dados); 5) *open science* @ Fapesp.

Partindo da premissa de que seria possível encontrar na página da Fapesp um modelo de PGD que atendesse de forma clara os cientistas da área de humanidades nessa função, acreditou-se que também estaria disponível uma definição, e até mesmo uma classificação, dos dados de pesquisa na área, o que não ocorreu.

De acordo com o site, o que se insere em um PGD vai variar conforme o campo disciplinar ao qual ele atenderá, além de levar em consideração como os responsáveis pela pesquisa pretendem disponibilizar os dados. Algumas chamadas para financiamento têm um formato específico (mas no site não está claro se essa exigência compreende áreas específicas ou financiamentos específicos), para as demais é exigido um texto de até duas páginas com as informações:

> a) Descrição dos dados e metadados produzidos pelo projeto – por exemplo, amostras, registros de coleta, formulários, modelos, resultados experimentais, software, gráficos, mapas, vídeos, planilhas, gravações de áudio, bancos de dados, material didático e outros.

1 Ver: http://www.fapesp.br/gestaodedados.

> b) Quando aplicável, restrições legais ou éticas para compartilhamento de tais dados, políticas para garantir a privacidade, confidencialidade, segurança, propriedade intelectual e outros.
> c) Política de preservação e compartilhamento (por exemplo, compartilhamento imediato ou apenas após a aceitação da publicação associada). Período de carência (antes do compartilhamento) e período durante o qual os dados serão preservados e disponibilizados.
> d) Descrição de mecanismos, formatos e padrões para armazenar tais itens de forma a torná-los acessíveis por terceiros. Esta descrição pode incluir o uso de repositórios e serviços de outras instituições. (Fapesp, s.d.)

Se o pesquisador da área de humanidades estiver pouco habituado com a linguagem relacionada a dados de pesquisa, é possível que a elaboração do PGD seja dificultosa mesmo quando estiver informado que dados são "amostras, registros de coleta, formulários, modelos, resultados experimentais, software, gráficos, mapas, vídeos, planilhas, gravações de áudio, bancos de dados, material didático *e outros*" (Fapesp, s.d., grifo nosso).

Para auxiliar o pesquisador, a Fapesp indica ainda dois links que direcionam para páginas estrangeiras onde estão disponibilizados modelos de PGD: um link associado a planos de gestão de dados nos EUA e outro em países do continente europeu. Uma busca pelo Google com o termo "como criar um plano de gestão de dados", selecionando nas ferramentas "em português" e "Brasil", mostrou que não temos, on-line, sites e ferramentas nacionais que auxiliem o pesquisador na montagem de seu plano de forma didática além da Agência USP de Gestão de Informação Acadêmica (Aguia),[2] que é um órgão da Universidade de São Paulo. Também temos a publicação *Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores* (Sayão; Sales, 2015), que passou a ser um dos materiais de referência no campo e sempre citado.

Uma reflexão sobre os resultados dessa busca aponta que o peso institucional com relação ao compromisso e à produção de conhecimentos sobre os PGD acaba caindo mais sobre os autores que se tornaram referência nesses estudos do que em planos institucionais. E estas instituições parecem estar buscando agora meios para inserir esses planos em suas políticas.

Além disso, não foi possível identificar uma definição para dados de pesquisa de maneira geral, ou dados de pesquisa em humanidades de maneira particular. No item da página da Fapesp intitulado "Documentos e páginas de interesse

2 Ver: https://www.aguia.usp.br/apoio-pesquisador/dados-pesquisa/plano-gestao-dados-2./

– Planos de gestão de dados", são disponibilizados alguns endereços nos quais podem ser consultados manuais, exemplos de planos e como redigi-los, para as mais variadas disciplinas.

A National Science Foundation (EUA), principal agência federal estadunidense de apoio à pesquisa, é um dos endereços informados que disponibiliza requisitos e planos de gestão de dados por área, com a seguinte classificação: *biological sciences* (BIO); *computer & information sciences & engineering* (Cise); *education & human resources* (EHR); *engineering* (ENG); *geosciences* (GEO); *mathematical and physical sciences* (MPS); e a área de interesse neste trabalho, *social, behavioral, and economic sciences* (SBE), que engloba antropologia; ciência cognitiva; ciência do desenvolvimento; economia; geografia; psicologia; políticas públicas; e sociologia (tradução nossa).

Ao acessar o documento disponibilizado para a área social, *behavioral and economic sciences* (SBE), intitulado *Data management for NSF SBE directorate proposals and awards* (NSF, 2018), tem-se a informação de que o documento se destina a oferecer orientação aos pesquisadores da área para o desenvolvimento de seus PGD. Mas também está disponibilizado nele o link para dois outros documentos, um deles intitulado *Public access to NSF: funded research data for the social, behavioral, and economic sciences* (NSF, 2016), um relatório que faz uma consideração cuidadosa e completa dos problemas de gerenciamento de dados para as ciências da área SBE, originário das discussões de um *workshop*. As discussões foram realizadas pelo grupo de trabalho formado com representantes das diferentes áreas do programa no SBE, que tinha como objetivo delinear "orientações mais detalhadas sobre os planos de gerenciamento de dados, para que as normas dos revisores e preparadores fossem articuladas e os dados se tornassem acessíveis e interoperáveis" (NSF, 2016, tradução nossa).

O que chama atenção nesse relatório é que ele, diferente dos documentos das outras áreas, e semelhante ao documento da área SBE, define o que é dado.

> Os dados são definidos para os fins deste documento como informações relevantes para, ou de interesse para pesquisadores sociais, comportamentais e econômicos, como entradas ou saídas de pesquisas. São materiais de pesquisa resultantes da coleta ou criação de dados primários ou derivados de fontes existentes. (NSF, 2016, tradução nossa)

É importante chamar a atenção para a diferença entre dados primários e dados derivados, principalmente com relação à forma como eles foram coletados. Dados primários são aqueles originais, coletados pelo pesquisador pela primeira vez, como os dados gerados pela aplicação de questionário, por exemplo.

Já os dados derivados são aqueles oriundos de fontes disponíveis, como análises realizadas pelo pesquisador que coletou os dados, com a aplicação de questionário, e que resultou em um artigo, por exemplo.

Também é interessante a frase "como entradas ou saídas de pesquisas", que nos remete à ideia de sistemas de informação, em que os dados alimentam o sistema [ENTRADA], são processados [PROCESSOS] e se transformam em informações [SAÍDA], mas essas informações também podem ser novos dados que retroalimentam o sistema, representado pela Figura 1:

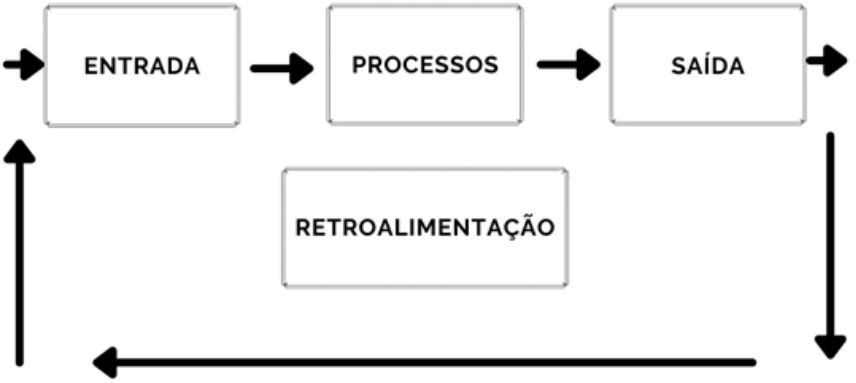

Figura 1 – Ciclo de processamento de dados. Fonte: elaborado pela autora

Já no *Data management for NSF SBE directorate proposals and awards*, documento da área *social, behavioral and economic sciences* (SBE), dado é definido pelo que ele é e pelo que ele não é.

> Os dados da pesquisa são definidos como o material factual registrado, comumente aceito na comunidade científica, conforme for necessário para validar os resultados da pesquisa, mas não são nenhum dos [materiais] seguintes: análises preliminares, rascunhos de trabalhos científicos, planos para pesquisas futuras, revisões por pares ou comunicações com colegas. Este material "gravado" exclui objetos físicos (por exemplo, amostras de laboratório). Os dados da pesquisa também não incluem:
> (a) segredos comerciais, informações comerciais, materiais necessários para serem mantidos em sigilo por um pesquisador até a publicação ou informações semelhantes protegidas por lei; e
> (b) informações pessoais e médicas e informações similares cuja divulgação constituiria uma invasão claramente injustificada da privacidade pessoal, como informações que poderiam ser usadas para identificar uma pessoa em particular em um estudo de pesquisa. (NSF, 2018, tradução nossa)

A publicação *Public access to NSF: funded research data for the social, behavioral, and economic sciences* (NSF, 2016) tem um apêndice em que relaciona os tipos de dados por área no item "Appendix E: example types of data used in the social, behavioral, and economic sciences", que contém uma lista extensa, mas que não se pretende exaustiva, dos tipos de dados identificados pelos participantes do *workshop* em seus resumos pré-*workshop*. Foram relacionados abaixo os tipos de dados identificados por área SBE, segundo o relatório.

Quadro 1 – Appendix E: example types of data used in the social, behavioral, and economic sciences

Campo	Exemplos de tipos de dados usados em SBE
Antropologia	Entrevistas gravadas e transcritas Notas de campo escritas Fotos Mapas: desenhados à mão ou digitais Gravações: áudio ou visual Dados quantitativos Fontes secundárias Meios de comunicação Texto interativo baseado na web Dados que foram analisados usando software qualitativo de análise de dados
Ciência cognitiva	Medidas de resultado Tempos de resposta Taxas de erro Gravações em áudio ou vídeo de respostas verbais ou motoras Medidas on-line Rastreamento ocular Rastreamento do *mouse* Potenciais relacionados ao evento Imagem cerebral funcional
Ciência do desenvolvimento	Gravações de vídeo ou áudio Transcrições de trocas verbais e comportamentais Questionários Dados baseados em computador Tela sensível ao toque Rastreamento ocular Tutores Arquivos simples baseados em texto para análise estatística
Economia	Experimental Observacional Pesquisa Conjuntos de dados vinculados Federal (por exemplo, microdados do censo) Dados corporativos (compartilhados apenas com um contrato de confidencialidade)
Geografia	Modelos Dados do Sistema de Informações Geográficas (SIG) Dados do Sistema de Posicionamento Global (GPS) Notas de campo Dados e imagens de sensoriamento remoto Dados administrativos

Psicologia	Gravações de áudio Gravações de vídeo Transcrições Pesquisas Testes padronizados Dados experimentais
Políticas públicas	Pesquisas Entrevistas Áudio Vídeo Métodos usados para geração de dados
Sociologia	Dados de pesquisa Arquivos contextuais construídos e mantidos por administradores de pesquisa Dados administrativos Entrevistas detalhadas Gravações Transcrições Notas etnográficas de campo Dados de biomarcadores coletados a partir de medições físicas Amostras biológicas dos entrevistados

Fonte: NSF, 2016, tradução nossa.

A lista reunida no Quadro 1 é a mais completa que encontramos até agora identificando, ao menos em parte, dados de pesquisa na área de humanidades. Nesse sentido, compreendemos que sua adoção, no todo ou parcial, como parâmetro para pensarmos o cenário brasileiro pode ser um bom percurso a se seguir no que tange à disseminação da questão referente aos dados de pesquisa em humanidades, sua gestão e disponibilização aos pares, visando à manutenção da pesquisa de cunho humanístico cada vez mais marcada pela sua própria transformação digital – e acarretando o aumento de práticas ligadas às humanidades digitais (Pimenta, 2020).

A pesquisa no Google usando o termo *"research data"* levou a diferentes páginas que definem dados de pesquisa de forma geral; a opção foi acessar somente os links que se encontravam na primeira página da pesquisa e remetiam a documentos oficiais de universidades.

No site da University of Leeds, os dados de pesquisa são definidos como "quaisquer informações que foram coletadas, observadas, geradas ou criadas para validar os resultados originais da pesquisa", e embora geralmente sejam digitais, "os dados de pesquisa também incluem formatos não digitais, como cadernos de laboratório e diários" (University of Leeds, s.d., tradução nossa).

Já no âmbito da University of Leicester, na página intitulada What is Research Data,[3] o pesquisador se depara com a informação inicial de que definir dados de pesquisa por si só já é um desafio, pois não existe consenso sobre sua definição, eles

3 Ver: https://www2.le.ac.uk/services/research-data/old-2019-12-11/rdm/what-is-rdm/research-data.

variam de acordo com a disciplina e com o financiador da pesquisa. Ao final dessa página, está disponível o link para o documento *Research data: definitions* (Burnham, 2012), que reúne "várias definições e explicações úteis" (tradução nossa), iniciando por: "Os dados da pesquisa, diferentemente de outros tipos de informações, são coletados, observados ou criados para fins de análise para produzir resultados originais da pesquisa" (Burnham, 2012, tradução nossa). O que chama atenção no texto é que ele aponta a dificuldade de definir dados de pesquisa, porque eles variam de acordo com a disciplina e o financiador, ou seja, como existe dificuldade de consenso, as definições podem ser feitas por área de pesquisa. Borgman (2012) pontua que a identificação do que são dados é subjetiva e está vinculada ao fazer acadêmico:

> Os dados podem existir apenas aos olhos de quem vê: o reconhecimento de que uma observação, artefato ou registro constitui dado é em si um ato acadêmico. Os curadores de dados, bibliotecários, arquivistas e outros envolvidos no gerenciamento de dados podem receber uma coleção que é considerada dado pelo coletor, mas não percebida como tal pelos destinatários. Por outro lado, um pesquisador pode estar segurando coleções de materiais sem perceber o quão valiosos eles podem ser como dados. *O conceito de dados é difícil de definir, pois os dados podem assumir muitas formas, tanto físicas quanto digitais*. Entre as definições mais amplamente citadas temos esta, de um relatório da National Academies of Science: "Os dados são fatos, números, letras e símbolos que descrevem um objeto, ideia, condição, situação ou outros fatores". (National Research Council, 1999, p. 15 apud Borgman, 2012, p. 3, grifo nosso)

A autora chama a atenção para o fato de a noção de dados ser menos desenvolvida nas humanidades, e ressalta que o crescimento das pesquisas em humanidades digitais levou ao uso mais comum do termo na área (Borgman, 2012, p. 3).

No documento *Research data: definitions* (Burnham, 2012), não apenas são encontradas diferentes definições de dados, mas também sua classificação a partir da forma como são gerados e os formatos em que podem ser encontrados. Com relação aos dados de pesquisa gerados na área das ciências humanas e sociais, somente três definições relacionadas à área são dispostas ao longo do documento:

> Todos os pesquisadores trabalham com dados, mas o que você chama de dados dependerá da sua disciplina. Como estudioso de ciências humanas, você pode falar sobre suas fontes ou textos principais. Se sua pesquisa é em ciências sociais, você pode pensar em termos de resultados de pesquisas, entrevistas e estatísticas. Você provavelmente terá termos diferentes novamente para os resultados de suas experiências e observações se for um cientista [das ciências naturais ou físicas]. (Burnham, 2012, tradução nossa)

A definição acima é da Monash University, a citação abaixo é de Neil Beagrie, Brian Lavoie e Matthew Woollard, do Keeping Research Data Safe 2 (KRDS2),[4] ambas presentes no documento da University of Leicester:

> Para os fins do estudo KRDS2, os dados da pesquisa são definidos como coletas de dados digitais estruturados de quaisquer disciplinas ou fontes que possam ser usadas por pesquisadores acadêmicos para realizar suas pesquisas ou fornecer um registro comprobatório delas. Os dados da pesquisa podem ser criados em vários contextos diferentes: por razões totalmente não relacionadas à pesquisa acadêmica; para pesquisa acadêmica ou como subproduto da pesquisa (acadêmica). Incluem uma grande variedade e heterogeneidade de dados e os respectivos metadados e documentação para torná-los utilizáveis e compreendidos, ou as representações e registros digitais para dados de pesquisa física. Em essência, qualquer tipo de dado de pesquisa já mantido em repositórios de dados estaria no escopo. Os exemplos podem incluir: dados complexos usados em modelagem climática, aerodinâmica, modelagem molecular, bioinformática; arquivos de vídeo e imagem utilizados em arqueologia, história da arte, antropologia e performance; imagens digitais/dados investigativos de fontes físicas primárias nas humanidades; dados quantitativos e qualitativos utilizados nas ciências sociais; ou dados e índices eletrônicos para amostras de fósseis ou tecidos da pele. (Burnham, 2012, tradução nossa)

E por fim, a definição abaixo, do Joint Information Systems Committee (JISC), também retirada do mesmo documento da University of Leicester:

> Pesquisadores em quase todas as disciplinas agora criam dados em formato digital. Esses dados podem vir de várias formas: por exemplo, as medições registradas por satélites de monitoramento ambiental, os produtos de colisões entre partículas fundamentais, as sequências de genomas inteiros, os resultados de pesquisas e entrevistas em ciências sociais, as imagens anotadas de inscrições gregas antigas ou os vídeos anotados de rotinas inovadoras de dança. (Burnham, 2012, tradução nossa)

No site da biblioteca da NC State University, na página Defining Research Data (uma adaptação da página elaborada pela biblioteca da University of Oregon), a definição de dados de pesquisa encontrada é:

4 Ver: https://beagrie.com/krds/.

> O material factual registrado geralmente aceito na comunidade científica como necessário para validar os resultados da pesquisa. (Circular OMB 110). Os dados de pesquisa abrangem uma ampla variedade de tipos de informações e os dados digitais podem ser estruturados e armazenados em uma variedade de formatos de arquivo. Observe que gerenciar dados (e registros) corretamente não equivale necessariamente a compartilhar ou publicar esses dados.[5]

Uma leitura nas outras páginas recuperadas na pesquisa possibilitou perceber que as definições de dados de pesquisa são muito abrangentes, e não foi encontrada nenhuma definição que se voltasse para os dados gerados nas pesquisas em humanidades especificamente.

Ou seja, a definição de dados de pesquisa na área das humanidades é essencial para a seleção de metadados e para a arquitetura dos repositórios, locais onde os conjuntos de dados serão depositados, mas não é possível considerar como consistentes e convincentes as que foram encontradas.

> Com relação aos metadados e documentação dos dados de pesquisa, Kindling e Schirmbacher (2013) apontam que são essenciais para permitir sua posterior utilização, pois descrevem o contexto de tais dados e as ferramentas com as quais eles foram gerados, armazenados, processados e analisados. (Vidotti et al., 2017, p. 8)

Para além da identificação do que são dados de pesquisa na área, um outro importante ponto a ser colocado diz respeito a sua gestão, que permite seu reúso e compartilhamento. Uma gestão correta e apropriada contribui para a reprodutibilidade da pesquisa, e o compartilhamento dos dados não apenas permite seu reúso, como também dá maior credibilidade aos produtos resultantes, como *papers*, artigos etc.

A área das humanidades no Brasil

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), órgão que coordena o Sistema Nacional de Pós-Graduação brasileiro e a formação inicial e continuada de professores para a educação básica, tem, dentre suas atividades, a avaliação da pós-graduação *stricto sensu* e o acesso e divulgação da produção científica, que são atividades avaliativas. Com a finalidade de facilitar as

5 Ver: https://www.lib.ncsu.edu/do/data-management/defining-research-data.

atividades de avaliação, as áreas (num total de 49) são agregadas, por critério de afinidade, em dois níveis: primeiro nível – colégios; segundo nível – grandes áreas. No Quadro 1 foram relacionados os campos que fazem parte, nos EUA, da área *social, behavioral, and economic sciences* (SBE).

De acordo com a classificação da Capes[6] (consulta feita no ano de 2021), as grandes áreas que são hierarquicamente relacionadas ao colégio de humanidades compreendem:

a) ciências humanas: antropologia/arqueologia; ciência política e relações internacionais; ciências da religião e teologia; educação; filosofia; geografia; história; psicologia; sociologia.

b) ciências sociais aplicadas: administração pública e de empresas, ciências contábeis e turismo; arquitetura, urbanismo e *design*; comunicação e informação; direito; economia; planejamento urbano e regional/demografia; serviço social;

c) linguística, letras e artes: artes; linguística e literatura.

Dentre essas grandes áreas, é possível perceber que a área *social, behavioral, and economic sciences*, que está sendo usada como referência na classificação de dados de pesquisa, mescla o que é identificado no Brasil como ciências humanas e ciências sociais aplicadas.

Fica nítido que a noção do que é dado de pesquisa sofre variações entre diferentes pesquisadores de uma mesma área e, mais ainda, entre áreas distintas do conhecimento. Sendo esses dados obtidos por diferentes processos, torna-se necessário o uso de tipologias específicas que deem conta de toda essa multiplicidade.

Como dito na introdução, este artigo propõe discutir o tema de dados de pesquisa na área das humanidades, buscando apresentar uma proposta de padronização na forma de uma classificação tipológica, com o intuito de auxiliar os pesquisadores da área na elaboração de seus planos de gestão de dados e no processo de depósito para preservação em repositórios, que certamente não pretende ser hermética.

A partir da publicação *Public access to NSF: funded research data for the social, behavioral, and economic sciences* (NSF, 2016), que relaciona os tipos de dados por área identificados pelos participantes de um *workshop* (vide Quadro 1), e de pesquisas realizadas no pós-doutorado, é apresentada uma classificação tipológica propondo-se lançar a pedra fundamental de um debate e de uma reflexão que se faz urgente. Diante do fato de que cada vez mais as tecnologias apresentam ao pesquisador novas ferramentas e possibilidades de pesquisa, muitas outras classificações de dados

6 https://www.capes.gov.br/avaliacao/sobre-as-areas-de-avaliacao.

podem vir a surgir, assim como outras podem não mais ser possíveis. O Quadro 2 reúne os tipos de dados a partir de uma divisão entre dados analógicos e digitais.

Quadro 2 – Proposta de uma classificação tipológica de dados de pesquisa em humanidades

Dados de pesquisa em formato não digital[7]	Dados de pesquisa em formato digital
Notas de campo escritas em cadernos de pesquisa Fotos (impressas) Mapas feitos à mão Transcrições de trocas verbais e comportamentais (impressas) Questionários (impressos) Experimental Observacional Modelos Transcrições (impressas) Testes padronizados (impressos) Métodos usados para geração de dados Dados de biomarcadores coletados a partir de medições físicas Amostras biológicas dos entrevistados	Notas de campo escritas em cadernos de pesquisa digitais ou digitalizadas posteriormente Fotos digitais ou digitalizadas posteriormente Mapas digitais ou digitalizados posteriormente Dados quantitativos Dados que foram analisados usando software qualitativo de análise de dados Medidas de resultado: tempos de resposta; taxas de erro Gravações em áudio ou vídeo de respostas verbais ou motoras Medidas on-line: potenciais relacionados ao evento; rastreamento ocular; rastreamento do *mouse* Imagem cerebral funcional Gravações de vídeo ou áudio Transcrições de trocas verbais e comportamentais Questionários Dados baseados em computador: tela sensível ao toque; rastreamento ocular; tutores Arquivos simples baseados em texto para análise estatística Conjuntos de dados vinculados Dados públicos Dados corporativos (compartilhados apenas com um contrato de confidencialidade) Modelos Dados do Sistema de Informações Geográficas (SIG) Dados do Sistema de Posicionamento Global (GPS) Dados e imagens de sensoriamento remoto Dados administrativos Transcrições Dados experimentais Material audiovisual Arquivos contextuais construídos e mantidos por administradores de pesquisa Dados de biomarcadores coletados a partir de medições físicas Imagens E-mail Memes Textos interativos on-line Blog Redes sociais em geral (Facebook, Twitter, Instagram, LinkedIn e outros) Artigos com arquivos anexados, geralmente oriundos de uma apresentação Códigos *Scripts* *Dataset* (CSV, planilha, Jason, Excel) Grafos

Fonte: elaborado pela autora com base em NSF, 2016.

7 Esses dados nascem em formato não digital, mas podem migrar para o formato digital quando são digitalizados, caso seja possível.

Embora muitos pesquisadores não tenham dimensão, eles passam todo o processo de suas pesquisas coletando/gerando, gerenciando e analisando dados para, posteriormente, publicar seus resultados. Hoje, como forma de preservação, vem se tornando uma prática comum o depósito de conjuntos de dados de pesquisa em repositórios criados para esse propósito. Para garantir a preservação desses dados, considerando que um *dataset* seja um objeto digital, é necessário seu depósito em um repositório, um local de armazenamento, de guarda e arquivamento de objetos digitais definidos como "um item armazenado em uma biblioteca digital, que consiste em dados, metadados e um identificador" (Arms, 2000, tradução nossa). Além da preservação, o depósito desses conjuntos de dados também se tornou uma exigência em diversos periódicos, como forma de comprovação dos resultados apresentados.

> Nas ciências físicas e da vida, a maioria dos dados é coletada ou produzida por pesquisadores, como observações, experimentos ou modelos. Nas ciências sociais, os pesquisadores podem coletar ou produzir seus próprios dados ou obter dados de outras fontes, como registros públicos de atividade econômica. A noção de dados é menos desenvolvida nas humanidades, embora o crescimento da pesquisa em humanidades digitais tenha levado ao uso mais comum do termo. Os dados de humanidades geralmente são extraídos de registros da cultura humana, sejam materiais de arquivo, documentos publicados ou artefatos. (Borgman, 2011, p. 6, tradução nossa)

Nielsen e Hjørland (2014, p. 225) vão citar a definição de Kaase (2001, p. 3251) para dados: "Dados são informações sobre propriedades de unidades de análise". Assim, para os autores, ao aceitarmos que "diferentes projetos de pesquisa possuirão diferentes unidades de análise", teremos uma definição ampla e relativa o suficiente para servir como uma definição geral de dados. "Está implícito nesta definição que os dados são sempre registrados com base em alguns interesses, perspectivas, tecnologias e práticas situadas que determinam seu significado e utilidade em diferentes contextos" (Nielsen; Hjorland, 2014, p. 225, tradução nossa).

Sendo assim, é possível inferir que cada área do conhecimento seria um contexto com interesses, perspectivas, tecnologia e práticas próprias, o que demandaria a necessidade de uma classificação de dados que pudesse atender cada uma delas. Não se está sugerindo que cada área terá tipos de dados exclusivos, mas afirmando que elas terão tipos de dados comuns a diferentes áreas e próprios de cada uma delas.

Conclusões

O pesquisador que está fazendo a gestão do ciclo de vida de seus dados de pesquisa precisa seguir algumas etapas, como, por exemplo, descrever os dados, utilizando metadados apropriados, garantir a preservação em repositório confiável e possibilitar que eles sejam recuperados, dentre outras. Essas três etapas foram selecionadas porque envolvem diretamente a classificação desses dados, pois essa ação no momento de seu arquivamento é essencial para possibilitar que eles sejam encontrados. Ainda que eles estejam acessíveis, ou seja, estejam disponíveis em acesso aberto em um repositório confiável, não terão valor se não puderem ser recuperados por outros pesquisadores.

A área das humanidades muitas vezes não é percebida como uma área que utiliza as tecnologias em seu processo de produção de conhecimento, ou reconhecida como uma área que gera dados em suas pesquisas, inclusive que poderão ser reutilizados em pesquisas de outras áreas, por isso é importante a sistematização desses dados.

O uso de uma classificação de dados em ciências humanas visa padronizar sua representação e, assim, facilitar seu acesso, busca, descoberta e recuperação, padronização essa que atenda o mais amplamente possível os diferentes tipos de dados gerados na área, mas que não se esgote. Mesmo os dados de pesquisa em formato não digital podem ser considerados objetos digitais, a partir de sua digitalização ou do depósito do *dataset* que o descreva.

Referências

ARMS, W. Y. *Digital libraries*. Cambridge, MA: MIT Press, 2000.

BORGMAN, C. L. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, p. 1-40, 2011. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1869155. Acesso em: 21 mar. 2020.

_____. Research data: who will share what, with whom, when, and why? *SSRN Electronic Journal*, september 9, 2010. Disponível em: https://www.researchgate.net/publication/48264305_Research_Data_Who_Will_Share_What_with_Whom_When_and_Why. Acesso em: 21 jul. 2020.

BURNHAM, A. *Research data*: definitions. University of Leicester, 4 set. 2012. Disponível em: https://www2.le.ac.uk/services/research-data/old-2019-12-11/documents/UoL_ReserchDataDefinitions_20120904.pdf. Acesso em: 21 mar. 2020.

EAVES, D. *The three laws of open government data*. Eaves.ca (blog), 30 set. 2009. Disponível em: https://eaves.ca/2009/09/30/three-law-of-open-government-data/. Acesso em: 21 mar. 2020.

FAPESP. Fundação de Amparo à Pesquisa do Estado de São Paulo. *Plano de Gestão de Dados*. Disponível em: http://www.fapesp.br/gestaodedados/. Acesso em: 21 mar. 2020.

FOX, Peter; HENDLER, James. eScience semântica: o significado codificado na próxima geração de

ciência digitalmente aprimorada. In: HEY, Tony; TRANSLEY, Stewart; TOLLE, Kristin (org.). *O quarto paradigma*: descobertas científicas na era da eScience. São Paulo: Oficina de Textos, 2011.

LAKATOS, E. M.; MARCONI, M. de A. *Fundamentos da metodologia científica*. 7. ed. São Paulo: Atlas, 2010.

LIMA, J. S. Gestão de dados de pesquisa no contexto da ciência aberta. *Informação em Pauta*, v. 5, n. 2, p. 212-214, 2020.

MEDEIROS, C. B. Gestão de dados científicos: da coleta à preservação. *SciELO em Perspectiva*, 2018. Disponível em: https://blog.scielo.org/blog/2018/06/22/gestao-de-dados-cientificos-da-coleta-a-preservacao/. Acesso: 20 mar. 2020.

NIELSEN, H. J.; HJORLAND, B. 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: 21 mar. 2020.

NSF. National Science Foundation. *Data management for NSF SBE directorate proposals and awards*. May 15, 2018. Disponível em: https://www.nsf.gov/sbe/DMP/SBE_DataMgmtPlanPolicy_RevisedApril2018.pdf. Acesso em: 21 mar. 2020.

_____. Public access to NSF: funded research data for the social, behavioral, and economic sciences. *Workshop Report*, May 17, 2016. Disponível em: https://www.nsf.gov/sbe/reports/Public_Access_NSF_Workshop_Report_Final_Briefs.pdf. Acesso em: 21 mar. 2020.

PIMENTA, R. M. Por que humanidades digitais na ciência da informação? Perspectivas pregressas e futuras de uma prática transdisciplinar comum. *Informação & Sociedade*, v. 30, n. 2, 2020. Disponível em: https://doi.org/10.22478/ufpb.1809-4783.2020v30n2.52122. Acesso em: 11 mar. 2022.

POIKOLA, A.; VILLUM, C.; DIETRICH, D. et al. What is open data? In: *Open data handbook*. 2017. Disponível em: http://opendatahandbook.org/guide/pt_BR/. Acesso em: 23 mar. 2020.

SALES, L.; CAVALCANTI, M. Seleção e avaliação de coleções de dados digitais de pesquisa: uma possível abordagem metodológica. *Informação & Tecnologia* (Itec), João Pessoa, v. 2, n. 2, p. 88-105, jul./dez. 2015. Disponível em: http://www.periodicos.ufpb.br/ojs/index.php/itec/article/view/34134. Acesso em: 20 maio 2020.

SANTOS, P. X.; CLINIO, A. A construção de uma política e o debate político para gestão e abertura de dados de pesquisa: o papel das instituições de ensino e pesquisa. In: ENCONTRO DA REDE SUDESTE DE REPOSITÓRIOS INSTITUCIONAIS, 1., 2019, Rio de Janeiro. *Anais*... Rio de Janeiro: Fiocruz/Ibict/UFRJ, 2019. 25 p. Disponível em: https://www.arca.fiocruz.br/handle/icict/33341. Acesso em: 11 mar. 2022.

SAYÃO, L. F.; SALES, L. F. Afinal, o que é dado de pesquisa? *Biblos*, v. 34, n. 2, 2020. (Dossiê Tecnologias de Informação e Comunicação no Contexto da Ciência da Informação). Disponível em: https://periodicos.furg.br/biblos/article/view/11875. Acesso em: 6 abr. 2022.

_____. *Dados de pesquisa*: quem ama cuida. Ilustração de Joana Sales Marques. Brasília, DF: Comissão Nacional de Energia Nuclear (Brasil); Ibict, 2019. Disponível em: https://livroaberto.ibict.br/bitstream/123456789/1083/2/cartilha%20dados%20de%20pesquisa.pdf. Acesso em: 11 mar. 2022.

_____. *Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores*. Rio de Janeiro: CNEN/IEN, 2015.

SETZER, V. W. Dado, informação, conhecimento e competência. In: _____. *Os meios eletrônicos e a educação*: uma visão alternativa. v. 10. São Paulo: Escrituras, 2001. (Coleção Ensaios Transversais; versão revista e ampliada em 25 de maio 2015). Disponível em: https://www.ime.usp.br/~vwsetzer/dado-info.html. Acesso em: 7 mar. 2020.

UNIVERSITY OF LEEDS. *Research data management explained*: what is research data? Disponível em: https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained. Acesso em: 21 mar. 2020.

VIDOTTI, S. A. B. G.; CONEGLIAN, C. S.; ROA-MARTÍNEZ, S. M.; ARAKAKI, F. A.; BRANDT, M. B.; FERREIRA, A. M. J. F. C. Repositório de dados de pesquisa para grupo de pesquisa: um projeto piloto. *Informação & Tecnologia*, v. 4, n. 2, p. 221-242, 2017. Disponível em: http://hdl.handle.net/20.500.11959/brapci/101623. Acesso em: 11 mar. 2022.

Recebido em 22/7/2021
Aprovado em 7/3/2022