Informação em Pauta
Fortaleza, CE v. 6 2021 ISSN 2525-3468
DOI: 10.36517/2525-3468.ip.v6i00.2021.68088.1-18

ARTIGO

DIRETRIZES PARA O ESTABELECIMENTO DE UM *CHECKLIST* PARA CURADORIA DE DADOS DE PESQUISA

GUIDELINES FOR ESTABLISHING A CHECKLIST FOR RESEARCH DATA CURATION

Samile Andrea de Souza Vanz[1]
Caterina Marta Groposo Pavão[2]
Sônia Elisa Caregnato[3]
Paula Caroline Schifino Jardim Passos[4]
Ana Maria Mielniczuk de Moura[5]
Eduardo Nunes Borges[6]
Rene Faustino Gabriel Junior[7]
Rafael Port da Rocha[8]

[1] Professora do Departamento de Ciências da Informação da Universidade Federal do Rio Grande do Sul. Doutora em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul.
E-mail: samilevanz@terra.com.br

[2] Professora do Departamento de Ciências da Informação da Universidade Federal do Rio Grande do Sul. Doutora em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul.
E-mail: caterina@cpd.ufrgs.br

[3] Professora do Departamento de Ciências da Informação da Universidade Federal do Rio Grande do Sul. Doutora em Information Studies pela University of Sheffield.
E-mail: sonia.caregnato@ufrgs.br

[4] Professora do Departamento de Comunicação da Universidade Federal do Rio Grande do Sul. Doutora em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul.
E-mail: paulacarolinejardim@gmail.com

[5] Professora do Departamento de Ciências da Informação da Universidade Federal do Rio Grande do Sul. Doutora em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul.
E-mail: ana.mmoura@uol.com.br

[6] Professor do Centro de Ciências Computacionais da Universidade Federal do Rio Grande. Doutor em Computação pela Universidade Federal do Rio Grande do Sul.
E-mail: eduardoborges@furg.br

[7] Professor do Departamento de Ciências da Informação da Universidade Federal do Rio Grande do Sul. Doutor em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho.
E-mail: renefgj@gmail.com

[8] Professor do Departamento de Ciências da Informação da Universidade Federal do Rio Grande do Sul. Doutor em Computação pela Universidade Federal do Rio Grande do Sul.
E-mail: rafael.rocha@ufrgs.br

ACESSO ABERTO

Conflito de interesses: Os autores declaram que não há conflito de interesses.

Financiamento: Rede Nacional de Pesquisa.

Declaração de Disponibilidade dos dados: Todos os dados relevantes estão disponíveis neste artigo.

Recebido em: 30 mar. 2021.
Aceito em: 21 set. 2021.
Publicado em: 26 out. 2021.

Como citar este artigo:

VANZ, Samile Andrea de Souza *et al*. Diretrizes para o estabelecimento de um *checklist* para curadoria de dados de pesquisa. **Informação em Pauta**, Fortaleza, v. 6, p. 1-18, 2021. DOI: https://doi.org/10.36517/2525-3468.ip.v6i00.2021.68088.1-18.

RESUMO

A curadoria de dados de pesquisa é um desafio para muitas instituições de pesquisa e profissionais responsáveis pela gestão de repositórios de dados. A comunidade científica vem aderindo às práticas de compartilhamento e reuso, e nesse sentido, cada vez mais é preciso decidir acerca da preservação de determinados conjuntos de dados em detrimento de outros. Este artigo objetiva discutir curadoria de dados de pesquisa e políticas de avaliação e de seleção de dados considerando-se o contexto de repositórios de dados de pesquisa nacionais. Apresenta resultados de uma pesquisa documental sobre requisitos para curadoria de dados, estabelecidos por repositórios ou instituições dedicadas à pesquisa sobre o tema. O artigo propõe algumas diretrizes para elaboração de checklist de critérios que podem ser aplicados aos repositórios brasileiros.

Palavras-chave: repositórios de dados de pesquisa; curadoria digital; preservação de dados.

ABSTRACT

Research data curation is a challenge for many research institutions and professionals responsible for managing data repositories. The scientific community has been adhering to the practices of sharing and reuse, and in this sense, it is increasingly necessary to decide on the preservation of certain data sets to the detriment of others. This paper aims to discuss research data curation, data evaluation and selection policies considering the context of national research data repositories. It presents the results of a documentary research on requirements for data curation, established by repositories or institutions dedicated to research on the subject. The paper proposes some guidelines for a checklist of criteria that can be applied to Brazilian repositories.

Keywords: research data repositories; digital curation; data preservation.

1 INTRODUÇÃO

O compartilhamento e o reuso de dados de pesquisa são práticas que vêm se consolidando há alguns anos entre a comunidade científica. Conforme apontado por Bell (2011), as leis do movimento planetário são exemplos antigos dessas práticas: elas foram formuladas por Kepler a partir do catálogo de anotações das observações astronômicas de Brahe, para quem Kepler trabalhava como assistente. O exemplo ilustra que, já no início do século 17, era possível estabelecer a divisão entre mineração e análise e dados experimentais, a divisão entre o cuidado no arquivamento de dados e a possibilidade da criação de teorias a partir do reuso de dados coletados por outro pesquisador.

Na atualidade, quando os dados deixaram de ser registrados em cadernos de anotações e passaram a ocupar arquivos e mídias cada vez mais sofisticados, o tema tem despertado a atenção de pesquisadores, editores de revistas, agências financiadoras e, ainda, das instituições onde as pesquisas se desenvolvem. Preservar esses dados, especialmente os coletados por pesquisadores individuais ou pequenos laboratórios, é um desafio para toda a comunidade científica, seja pelo grande volume de dados produzidos, seja pela rápida obsolescência de softwares e de hardware para acesso a esses. Os dados gerados pelas atividades de pesquisa necessitam de cuidados específicos, demandando um modelo de gestão que inclua ações de arquivamento seguro, ações de preservação, formas de acrescentar valor aos conteúdos e garantias de reuso (SAYÃO; SALES, 2012), facilitando a busca e a descoberta.

Nessa perspectiva, entende-se que as práticas de compartilhamento e de reuso requerem duas frentes de trabalho. Primeiramente, a infraestrutura tecnológica necessária para garantir o ambiente para guarda dos dados - os repositórios. Em seguida, o planejamento, com a definição das políticas de seleção, preservação, acesso e reuso desses dados (SIEBRA et al., 2013). Um aspecto crítico foi colocado por Heumüller et al. (2020), a disponibilização de ferramentas para acesso e para reuso dos dados, o que facilita a replicação, a reprodução, a extensão, a avaliação e a verificação dos resultados. Nesse âmbito, estão as atividades de curadoria.

A curadoria digital é compreendida, de acordo com as ideias de Pennock (2007, p. 1, tradução nossa), como "manter e agregar valor a um corpo confiável de informação digital para uso atual e futuro: em outras palavras, é o gerenciamento ativo e a avaliação das informações digitais durante todo o seu ciclo de vida." A curadoria digital, portanto, está associada ao ciclo de vida do objeto, o que justifica ações que podem ser presentes ao longo de toda a sua vida, conforme modelo do Digital Curation Center (DCC). O DCC especifica ações para o ciclo de vida da preservação e para curadoria digital, classificando-as em ações sequenciais, ações ocasionais e ações envolventes (sempre presentes, independentemente da fase).

Entende-se que curadoria digital, preservação digital e acesso relacionam-se por serem relevantes durante todo o ciclo de vida digital. Além disso, são atividades complementares, embora consideradas díspares até pouco tempo: a preservação é destinada a fornecer acesso ao conteúdo gerenciado ao longo do tempo, enquanto o

acesso depende da preservação em um determinado momento (CALIFORNIA DIGITAL LIBRARY, 2010).

A preservação de dados ao longo do tempo requer repositórios digitais com governança sustentável e estrutura organizacional, infraestrutura confiável e políticas abrangentes que deem apoio às práticas acordadas pela comunidade (LIN et al., 2020). Sayão e Sales (2013) argumentam que, pela importância estratégica para a pesquisa científica, as atividades de gestão dos dados, compartilhamento e reuso devem estar apoiadas em compromissos institucionais de longo prazo. Ao se pensar sobre as práticas científicas e sobre as políticas dos repositórios, a decisão acerca do que manter em termos de dados de pesquisa tem sido reconhecida há algum tempo. O gerenciamento de dados de pesquisa e, em particular, a avaliação e a seleção vêm recebendo atenção nos últimos anos, à medida que crescem o seu volume e a sua diversidade (BEAGRIE, 2019).

Nesse contexto, algumas questões precisam ser respondidas: todos os conjuntos de dados devem ser guardados para sempre? Como saber o valor do dado tendo em vista a linha do tempo? Dados poderiam ser preservados por períodos determinados? Quais seriam os valores para determinar prazos de guarda? Os dados são um registro vital de um determinado projeto vinculado a um consórcio entre instituições e por isso devem ser preservados indefinidamente? Os conjuntos de dados servem e estão preparados para o reuso e/ou para reprodutibilidade de experimentos ou análises? Os dados depositados possuem qualidade suficiente para atender às necessidades de reuso? Como garantir qualidade ao dado? Preparar o dado para o compartilhamento é um processo caro, os dados são suficientemente preparados para a replicação de um estudo? Tenho os direitos legais e de propriedade intelectual para manter e reutilizar esses dados? Os dados podem ser negociados? As informações descritivas, como os metadados, são suficientes para que os dados sejam encontrados, independentemente de onde forem armazenados? Os metadados seguem um padrão aberto, bem estabelecido e interoperável, ou seja, permitem que diferentes sistemas computacionais heterogêneos localizem, indexem e processem os conjuntos de dados? Existem recursos financeiros e infraestrutura para manter o armazenamento dos dados por tempo determinado ou indefinidamente?

O presente artigo é decorrente de resultados de pesquisa desenvolvida em um esforço conjunto entre a Rede Nacional de Ensino e Pesquisa (RNP), a Universidade Federal do Rio Grande do Sul (UFRGS) e a Universidade Federal do Rio Grande (FURG),

que visa desenvolver estudos sobre compartilhamento dos dados coletados, gerados e utilizados pelos pesquisadores brasileiros. O artigo discute curadoria de dados de pesquisa e políticas de avaliação e de seleção de dados considerando-se o contexto de repositórios de dados de pesquisa nacionais. O estudo se fundamenta em uma metodologia de pesquisa documental, feita com base nas diretrizes e modelos dos seguintes repositórios de dados internacionais e relatórios de instituições da área: *Digital Curation Centre and Australian National Data Service Guide*; *UK Data Service Collection Development Selection and Appraisal Criteria*; *NERC Data Value Checklist; DCC Checklist of Appraising Research Data*; *Cambridge PrePARe Project*; *University of Bristol Research Data Evaluation Guide*. Como resultado da pesquisa documental, propõem-se algumas diretrizes em um checklist para avaliação e para seleção de dados de pesquisa aplicado à realidade brasileira.

Entende-se que o estabelecimento de políticas para seleção e para avaliação de dados de pesquisa pode apoiar os repositórios nacionais no sentido de orientar e justificar as decisões de aquisição de infraestrutura computacional em geral, seja hardware, contratação de serviços em nuvem, ou software, bem como contribuir para agilidade das atividades práticas relacionadas à coleta de dados e estimar os recursos humanos necessários para gestão dos dados. Ademais, o estabelecimento de diretrizes nesta área pode auxiliar no treinamento de profissionais em seleção e em avaliação de dados e direcionar no desenvolvimento de novos repositórios. As seções seguintes apresentam uma revisão sobre avaliação e seleção de dados de pesquisa e, em seguida, abordam a definição de políticas para essas atividades. Ao final, apresenta-se uma proposta de itens a serem considerados na avaliação e seleção de dados de pesquisa para repositórios brasileiros.

2 AVALIAÇÃO E SELEÇÃO DE DADOS DE PESQUISA

A “avaliação” é um termo técnico muito utilizado por arquivistas, todavia, frequentemente referido por profissionais de outras áreas como “seleção” ou “aquisição”. A atividade está intimamente ligada a um repositório, ou a uma política institucional de desenvolvimento de coleções. A avaliação é a função mais nobre, o núcleo central da prática arquivística contemporânea (WHYTE; WILSON, 2010).

Couture (2005) também cita a avaliação como uma das funções mais importantes da prática arquivística contemporânea. Para ela, as decisões resultantes da avaliação afetam outras atividades como o recolhimento, o arranjo, a descrição, a acessibilidade e a preservação dos documentos. Conforme a autora, a avaliação pode ser considerada como o ato de julgar os valores primários e secundários de documentos e estabelecer o período no qual eles retêm este valor, em um contexto que respeite as relações essenciais entre uma dada instituição (ou pessoa) e os documentos que eles criaram no curso de suas atividades (COUTURE, 2005, p. 83, tradução nossa). O National Archives (2013, p. 3, tradução nossa) define avaliação como "o processo de distinguir registros de valor contínuo daqueles sem valor adicional, para que este último possa ser eliminado".

No ciclo de vida da curadoria digital do DCC, avaliação e seleção compreendem o processo de avaliar os materiais, para decidir quais manter a longo prazo, quais manter por um tempo, e quais descartar (HIGGINS, 2012). Avaliar e selecionar dados de pesquisa envolve adaptar para o mundo digital práticas de avaliação e de seleção já desenvolvidas para acervos analógicos em arquivos e bibliotecas:

> Existe um consenso considerável sobre os critérios utilizados na avaliação e seleção de materiais não digitais. A prática de avaliação de arquivo usa o conceito de valor de arquivo, que é obtido considerando-se o valor administrativo (utilidade para a condução dos negócios), valor fiscal (utilidade para negócios financeiros), valor legal (digno para a realização de negócios legais), valor intrínseco (natureza inerente e significância do artefato), valor probatório (valor como registro das origens, funções e atividades do criador do registro) e valor informativo (utilidade do conteúdo para fins de pesquisa mais gerais) (TIBBO, 2003, p. 29-30[i], *apud* HARVEY, 2007, p. 4).

De acordo com Harvey (2007), a seleção de materiais para preservação a longo prazo em bibliotecas concentra-se na manutenção de itens físicos em seus formatos originais e aplica cinco critérios principais: valor probatório, valor estético, valor de mercado, valor associativo e valor da exposição. O autor comenta que critérios adicionais podem ser aplicados, como condição física, recursos disponíveis, uso e significado social.

A possibilidade de armazenamento com custos mais baixos não diminuiu a importância das atividades de avaliação e de seleção. Gerir os dados é mais custoso do que armazená-los, em um cenário atual em que o volume de dados produzido cresce exponencialmente. Segundo Higgins (2012), é uma falácia que todo o material digital possa ser mantido sem a necessidade de avaliação porque o armazenamento é barato, e

continua a se tornar cada vez mais barato. O autor pontua cinco razões principais para facilitar a avaliação dos dados da pesquisa:

a) reduzir a quantidade de material que deve ser gerenciado ou custodiado a longo prazo, permitindo que os recursos possam ser direcionados para os materiais que possuem valor a longo prazo;
b) facilitar a capacidade de manter o acesso intelectual ao material, criar e anexar metadados, indexar e armazenar logicamente, para que os dados possam ser pesquisados e recuperados com rapidez e eficiência;
c) garantir que atividades de preservação possam ser realizadas de uma maneira simples e organizada para assegurar melhor a longevidade dos dados;
d) limitar os custos de armazenamento e manuseio do material;
e) garantir que as obrigações legais de armazenamento e acesso a dados sejam cumpridas.

A avaliação é o processo pelo qual alguns registros são selecionados para preservação, enquanto outros (grande maioria) são considerados de valor insuficiente para justificar a guarda permanente. A seleção deve ser guiada por políticas locais e comunitárias e por requisitos. O processo usado para decisões de seleção deve ser transparente e responsável. Destarte, não pode ser baseado em visualizações individuais sobre potenciais necessidades de pesquisa. À vista disso, comunidades de pesquisa e instituições precisam desenvolver e concordar com um conjunto de critérios objetivos para avaliar o significado dos conjuntos de dados de pesquisa a longo prazo.

Baseados na categorização da National Science Foundation (NSF) para dados de pesquisa – observacionais, computacionais e experimentais – Sayão e Sales (2013) lançam algumas reflexões. Para os autores, dados observacionais, associados a lugares e a tempos específicos, caracterizam-se por sua natureza instantânea e, portanto, são registros históricos que não podem ser coletados em outro momento. São dados que devem ser preservados para sempre. Por outro lado, dados computacionais resultantes de simulações podem ser replicados em outros momentos, salvo casos em que há uma dependência de hardware e software. A preservação deste tipo de dado pode ser conveniente em alguns casos. Da mesma forma, dados experimentais provenientes de situações controladas em bancada de laboratório são, em tese, reproduzíveis. Os autores

ponderam, no entanto, o custo da reprodutibilidade de um experimento. O balanço entre o custo da curadoria dos dados e o custo da reprodução do experimento é um item a ser pensado por uma política de preservação.

3 POLÍTICAS DE AVALIAÇÃO E DE SELEÇÃO

Diversas instituições adotam políticas para a gestão dos dados de pesquisa de sua produção ou sob sua custódia. No entanto, poucos desses planos especificam os parâmetros para estimar o valor dos conjuntos de dados e as ações a serem adotadas no sentido de definir seus prazos de preservação, de reavaliação ou de descarte, quando adequado.

Assim, a especificação de uma política para avaliação e para seleção de dados de pesquisa tem se mostrado uma iniciativa necessária, mesmo frente à perspectiva tecnológica que aponta para a capacidade de armazenamento cada vez maior a um custo decrescente, o que permitiria manter todos os dados indefinidamente. Harvey (2007) adverte que essa abordagem se limita à preservação dos bits e que não é apropriada à preservação do conhecimento. Ele esclarece que não é prático, ou mesmo desejável, manter o acesso a todos os dados permanentemente e que é necessário manter um conjunto de dados com base em seu significado e valor permanente, para que se possa compreendê-los no futuro.

Além disso, o volume dos dados de pesquisa implica uma tarefa de curadoria insustentável. Com os instrumentos científicos recentemente desenvolvidos e com o crescente uso de simulações por computador, uma equipe de pesquisa pode gerar muitos terabytes de dados por dia. Os curadores de dados enfrentam o gerenciamento na magnitude de petabytes e muito além (NDSA Agenda Working Group, 2020). Sayão e Sales (2012) argumentam que o ‘dilúvio de dados’ é desencadeado principalmente pelo avanço de instrumentos, sensores e escalas, que aumentaram exponencialmente a capacidade de obtenção de dados pela realização de observações e de medições de fenômenos, somados às informações geradas artificialmente por simulações e por software. Na definição de políticas para curadoria, deve-se considerar o volume, a variedade e a velocidade com que os dados são gerados (SAYÃO; SALES, 2013). Estes três fatores são determinantes para o gerenciamento de Big Data, sendo ainda

importantes considerar a veracidade (qualidade e confiabilidade) dos dados e o valor agregado.

Conforme Siebra (2020), os objetivos da instituição responsável pelo repositório de dados, as políticas institucionais, os recursos e o tempo disponíveis, as especificidades do acervo e do público usuário são detalhes fundamentais ao planejamento de um modelo de curadoria. A autora argumenta que é preciso especificar por que a instituição deseja preservar os dados, além de conhecer as políticas e as normativas já existentes que possam afetar o planejamento do projeto de curadoria. Também é muito importante conhecer a comunidade que fará uso do repositório para depósito e/ou para reuso dos dados, considerando as características peculiares de cada área do conhecimento. Nesse sentido, a curadoria digital engloba ações gerenciais, técnicas, tecnológicas e políticas (SIEBRA, 2020), que necessitam um planejamento dividido nas seguintes dimensões: infraestrutura (física e tecnológica), recursos financeiros, recursos humanos, ações de preservação digital, questões legais e éticas, acesso e uso dos objetos digitais.

Whyte e Wilson (2010) especificaram sete critérios gerais que embasam a derivação de critérios mais específicos em uma política de avaliação e de seleção de dados de pesquisa. São eles:

a) Relevância: o conteúdo do recurso atende às funções da instituição e cumpre os requisitos estabelecidos pela instituição de pesquisa ou pelo órgão de financiamento, incluindo qualquer aspecto legal para preservação dos dados além de seu uso imediato;
b) Valor científico ou histórico: os dados são científica, social ou culturalmente importantes, ou seja, pode-se prever o uso futuro pelo valor da pesquisa;
c) Exclusividade: o recurso é a única ou a mais completa fonte de informação sobre o tema e está em risco de perda se não for aceito;
d) Potencial de redistribuição: a confiabilidade, a integridade e a usabilidade dos arquivos de dados podem ser determinadas, estes são recebidos em formatos que atendem aos critérios técnicos especificados e propriedade intelectual ou questões éticas são abordadas;
e) Não replicabilidade: não seria viável replicar os dados ou a replicação seria financeiramente inviável;

f) Viabilidade econômica: os custos para gerenciar e preservar o recurso podem ser estimados e são justificáveis a partir de evidências de possíveis benefícios futuros, além disso, o financiamento está previsto, quando apropriado;
g) Documentação completa: as informações necessárias para facilitar a descoberta, acesso e reutilização são abrangentes e corretas; incluindo metadados sobre a proveniência do recurso e o contexto de sua criação e uso.

O UK Data Service, conforme consta no documento intitulado *Collection Development Selection and Appraisal Criteria* (UK Data Service, 2018), apresenta critérios semelhantes aos postulados por Whyte e Wilson (2010), acrescentando, entretanto, os seguintes itens:

a) Novidade: os dados contêm informações que foram solicitadas pelos usuários ou representam lacuna na coleta atual;
b) Valor internacional: os dados contêm informações de interesse da comunidade científica internacional;
c) Necessidade de replicação: os dados e os recursos representam resultados necessários para replicar ou revisar as pesquisas.

Além dos critérios citados, pode ser acrescido pelo menos mais um, anteriormente identificado por Harvey (2007), a vulnerabilidade. Ela é determinada pela necessidade de medidas especiais para ler ou para acessar os dados ou pelas condições e idade da mídia na qual eles se encontram. Segundo Eaker (2016), os curadores do repositório precisam determinar se os custos adicionais para oferecer acesso diferenciado compensam os benefícios de ofertar aqueles dados. O UK Data Service (2018) também apresenta critérios para não publicar os dados, que são os seguintes: a existência de questões legais e éticas como direitos autorais e proteção que impediriam o uso pleno dos dados; a falta de materiais contextuais que permitam a reutilização dos dados; e a aplicação de formatos antigos, que tornam as informações ilegíveis ou de difícil conversão e recuperação.

Além de atender a critérios específicos para cada contexto, os dados de pesquisa poderão ser classificados em categorias que irão definir o tipo de curadoria conforme o tempo e o modo de disponibilização. Abaixo, apresentam-se cinco categorias de

curadoria empregadas nas coleções de dados do UK Data Service, conforme consta no documento intitulado *Collection Development Selection and Appraisal Criteria* (UK Data Service, 2018):

a) conjuntos de dados selecionados para curadoria de longo prazo e disponibilizados para acesso on-line ou download;
b) conjuntos de dados selecionados para gerenciamento de curto prazo. Mesmo não sendo selecionados para preservação por longo prazo, eles receberão cópias de segurança (preservação em nível de bit), serão disponibilizados e poderão ser recuperados por meio de ferramentas de acesso on-line ou em software de repositório interno;
c) conjuntos de dados selecionados apenas para "entrega" aos usuários finais por meio de uma interface do repositório, consistindo, por exemplo, em dados de terceiros acessados por APIs / serviços da Web;
d) conjuntos de dados selecionados apenas para “descoberta”, ou seja, são dados de terceiros que não serão formalmente incluídos no repositório, mas serão passíveis de recuperação via registro de metadados no repositório, permitindo que os dados sejam encontrados mais facilmente;
e) conjuntos de dados preservados que não estão no escopo do repositório. Os dados classificados nesta categoria poderão ser movidos para outras categorias, se necessário.

Os critérios acima elencados são norteadores, mas as políticas são implementadas de acordo com as necessidades locais. É necessário também considerar as partes envolvidas nos processos de avaliação e de seleção, assim como as necessidades dos produtores e dos consumidores, e estabelecer as responsabilidades dos profissionais e das instituições.

Em suma, uma política de avaliação e de seleção é fundamental para se ter um entendimento claro dos tipos de conjuntos de dados que podem fazer parte do repositório, dirimindo incertezas e estabelecendo limites e responsabilidades das partes envolvidas. Essa clareza ajuda também na determinação de como promover os serviços do repositório, identificando onde concentrar os esforços de divulgação (EAKER, 2016).

No entanto, mesmo políticas claras apresentam limites. Elas descrevem pontos-chave, por isso precisam ser complementadas por critérios específicos e procedimentos para sua implementação. Neste sentido, *checklists* são ferramentas que permitem operacionalizar as políticas, como tratado na próxima seção.

4 SELEÇÃO E AVALIAÇÃO EM REPOSITÓRIOS DE DADOS BRASILEIROS: PROPOSTA DE DIRETRIZES PARA *CHECKLIST*

Um dos elementos fundamentais acerca da decisão de disponibilização e preservação de dados de pesquisa é a exigência por parte dos financiadores da pesquisa. No Brasil, as agências de fomento têm implementado recomendações e algumas condições.

O CNPq informa em seus editais a responsabilidade dos pesquisadores, de suas equipes e de suas instituições em manter, sempre que possível, os resultados da pesquisa, dados e as coleções à disposição de outros pesquisadores para fins acadêmicos. Os artigos científicos resultantes dos projetos apoiados pela agência deverão ser publicados, preferencialmente, em periódicos de acesso público e depositados, em conjunto com os dados científicos e com todo material suplementar relacionado, em repositórios eletrônicos de acesso público (CONSELHO..., 2020).

Na FAPESP, o Plano de Gestão de Dados vem se tornando um componente obrigatório na fase de submissão de um projeto. Para determinadas modalidades e chamadas, o documento "Plano de Gestão de Dados" faz parte dos anexos obrigatórios de uma proposta submetida. O Plano reúne um conjunto de informações básicas – quais dados serão produzidos pelo projeto, restrições de compartilhamento, como serão compartilhados e como serão preservados (FAPESP, 2020). Ainda, orienta a redação de um texto de até duas páginas, contendo as seguintes 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;

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. Essa descrição pode incluir o uso de repositórios e serviços de outras instituições.

As informações disponibilizadas pelos pesquisadores em seus planos de gestão de dados de pesquisa são relevantes para a curadoria nos repositórios de dados, tendo em vista que respondem a várias perguntas que o profissional responsável pelo repositório precisa averiguar. Nesse sentido, considera-se bastante pertinente e atual a exigência para que os pesquisadores apresentem planos, por parte de agências de financiamento ou instituições de pesquisa.

Considerando-se a relevância de discutir critérios para curadoria, realizou-se uma pesquisa documental em diversos repositórios e grupos de pesquisa ligados ao tema. As diretrizes apresentadas no *ckecklist* do Quadro 1 tem por base os modelos do *Digital Curation Centre and Australian National Data Service Guide*, do *UK Data Service Collection Development Selection and Appraisal Criteria*, do *NERC Data Value Checklist*, o *DCC Checklist of Appraising Research Data*, do *Cambridge PrePARe Project*, e do *University of Bristol Research Data Evaluation Guide*. O *checklist* está organizado em critérios obrigatórios e critérios importantes. Orienta-se que, quando houver resposta positiva a pelo menos um dos critérios obrigatórios, os dados sejam selecionados para preservação. Em relação aos critérios importantes, orienta-se que quando for possível responder "Sim" a pelo menos uma das perguntas de cada seção os dados provavelmente sejam encaminhados para seleção para preservação. Os demais critérios são considerados de suporte (itens origem, condições, requisitos de armazenamento e preservação, limitações de acesso, limitações técnicas). Nesses casos, a seleção e a preservação ocorrerão caso seja possível responder "Sim" à maioria das perguntas.

Quadro 1 - Diretrizes para estabelecimento de um c*hecklist* para seleção e avaliação de dados de pesquisa aplicado à realidade brasileira.

Critérios obrigatórios Responder "Sim" a qualquer uma das perguntas abaixo resulta automaticamente na seleção para preservação		
Considerações legais / estatutárias	**Sim**	**Não**
Existe uma razão legal para reter os dados?		
Existe alguma razão para acreditar que os dados possam ser usados em litígios, inquéritos públicos, investigações policiais, ou qualquer relatório ou artigo que possa ser legalmente contestado?		
Os dados são o produto do financiamento de agências públicas e sustentam uma produção de pesquisa publicada?		
Existem obrigações contratuais que exigem a preservação dos dados?		
Políticas	**Sim**	**Não**
A política de Dados de Pesquisa do financiador de pesquisa exige que os dados sejam mantidos?		
Os dados serão citados em uma publicação com uma política que exige que os dados sejam disponibilizados?		
Aplica-se alguma orientação específica da área do conhecimento que solicita a preservação dos dados?		
Critérios importantes Responder "Sim" a pelo menos uma das perguntas de cada seção abaixo provavelmente deve resultar em uma seleção para preservação.		
Valor de reutilização	**Sim**	**Não**
O arquivo não é uma cópia de dados disponíveis em um repositório. Os dados são únicos e / ou impossíveis de reproduzir, como por exemplo, registram eventos que não podem ser recriados?		
Os dados têm apelo amplo e é provável que sejam do interesse de outras pessoas (por exemplo, uma ampla faixa geográfica ou temporal ou um foco interdisciplinar)?		
É provável que os dados tenham valor acadêmico especial (por exemplo, representam uma descoberta histórica) ou estabelecem um novo precedente importante que provavelmente será seguido por outros (por exemplo, envolve uma nova técnica de processamento de dados)?		

Contexto da pesquisa	**Sim**	**Não**
É provável que os dados sejam citados e referenciados em uma publicação acadêmica?		
Os dados agregam valor a alguma coleta de dados significativa, contribuindo para uma coleção preexistente?		
Os dados fundamentam novos projetos de pesquisa?		
Os dados têm potencial para reuso e citação em futuros projetos de pesquisa?		
Critérios de suporte Responder "Sim" à maioria das perguntas abaixo deve resultar em seleção para preservação.		
Origem	**Sim**	**Não**
A reprodução dos dados é cara ou difícil?		
Os dados têm sua integridade original? (por exemplo, não são processados e foram armazenados com segurança desde que foram gerados)		
O arquivo é a cópia de referência (definitiva) dos dados?		
Condições	**Sim**	**Não**
Os dados possuem metadados suficientes? (por exemplo, uma descrição no nível do catálogo, uma descrição de como os dados são organizados, documentação de como e por que os dados foram criados e um guia sobre como usá-los)		
Os dados são de qualidade adequada para depósito em um *Data Center* ou outro repositório? (ou seja, os dados são controlados pela qualidade, bem organizados, legíveis e íntegros)		
Há disponibilidade de dados equivalentes e mais relevantes?		
Requisitos de armazenamento e preservação	**Sim**	**Não**
Os dados podem ser armazenados sem exigências específicas		
Os dados podem ser preservados de uma forma utilizável (isto é, permanecem adequados à finalidade) sem quaisquer requisitos excepcionais?		

Existe financiamento para a preservação desses dados em particular (pela equipe de pesquisa, por uma instituição anfitriã ou por um *Data Center*)?		
Limitações de acesso	**Sim**	**Não**
Se houver dados pessoais envolvidos, foi obtido o consentimento informado dos sujeitos da pesquisa para arquivamento e reutilização de dados?		
Se for necessária a aprovação de um Comitê de Ética, há evidências de que este procedimento foi seguido?		
A natureza dos dados sugere restrições ao compartilhamento, acesso e reutilização? (por exemplo, conjunto de dados envolve dados políticos ou de saúde confidenciais)		
Os dados estão livres de quaisquer termos e condições que limitariam o acesso? (por exemplo, requisitos de licença de *software*, acordos comerciais que proíbem a reutilização)		
Limitações técnicas	**Sim**	**Não**
Os dados estão em um formato técnico aceitável, preferencialmente aberto, para depósito em um *Data Center*?		
Os dados são utilizáveis sem necessidade de nenhum software / hardware específico?		
Se "Não" à pergunta acima, o software / hardware específico está prontamente disponível?		
É possível gerar versões diferentes dos dados para aumentar o valor de reutilização (por exemplo, criar formatos de arquivo alternativos)?		

Fonte: Elaborado pelos autores, com base nos checklists do *Digital Curation Centre and Australian National Data Service Guide, UK Data Service Collection Development Selection and Appraisal Criteria, NERC Data Value Checklist, DCC Checklist of Appraising Research Data, Cambridge PrePARe Project, University of Bristol Research Data Evaluation Guide*.

A aplicação desse *checklist* deverá oferecer uma base sólida para a decisão da equipe responsável pela curadoria de dados nos repositórios de dados abertos.

5 CONSIDERAÇÕES FINAIS

As instituições de pesquisa e universidades têm trabalhado na criação de normativas para seus pesquisadores e para seus repositórios para depósito e para preservação dos dados de pesquisa. O desenvolvimento desses repositórios é um processo técnico, no entanto, seu uso envolve decisões importantes sobre gestão, guarda e preservação dos dados no repositório. Para apoiar estas decisões, diversas instituições internacionais têm disponibilizado aos seus pesquisadores e gestores *checklists* para

seleção e avaliação de dados, de modo a facilitar a interpretação e a definição acerca da guarda de um determinado conjunto de dados. Nesse sentido se propõe este *checklist,* baseado em documentos internacionais, que identifica requisitos importantes ao contexto de pesquisa brasileiro.

Entre os *checklists* analisados e que serviram de base para a presente proposta estão os do *Digital Curation Centre and Australian National Data Service Guide*, do *UK Data Service Collection Development Selection and Appraisal Criteria*, do *NERC Data Value Checklist*, do *DCC Checklist of Appraising Research Data*, do *Cambridge PrePARe Project*, e do *University of Bristol Research Data Evaluation Guide*. Os documentos internacionais direcionaram a proposta que ora se apresenta, que sugere-se, seja avaliada após ser utilizada por repositórios brasileiros temáticos, multidisciplinares e institucionais.

Os dados gerados pelos pesquisadores devem ser preservados por cientistas de dados, bibliotecários, arquivistas e programadores, profissionais que precisam trabalhar em total sintonia com a comunidade científica nacional e internacional. É necessário conhecer o dado desde sua criação e curá-lo para além do término do projeto de pesquisa que o gerou, trabalhar na integração de dados que são em geral, fragmentados, espalhados em nível mundial. A curadoria de dados é um mecanismo para garantir a confiança de quem vai depositar e daqueles que vão utilizar o dado, por isso ela precisa ser pensada em nível internacional. O mundo inteiro se debruça sobre a implementação de técnicas e políticas para curadoria, por isso é fundamental a discussão do tema no Brasil. Para que os dados de hoje possam servir para o desenvolvimento de teorias como as desenvolvidas por Kepler no início do século 17.

REFERÊNCIAS

BEAGRIE, Neil. **What to keep**: a JISC research data study. 2019. 64 p.

BELL, Gordon. Prefácio. *In*: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (org.) **O quarto paradigma**: descobertas científicas na era da eScience. São Paulo: Oficina do Texto, 2011.

CALIFORNIA DIGITAL LIBRARY. UC3Curation Foundations. UC Curation Center / California Digital Library, 2010. Disponível em: https://confluence.ucop.edu/download/attachments/13860983/UC3-Foundations-latest.pdf?version=1. Acesso em: 08 jun. 2020.

COLE, Gareth; EVANS, Jill; LLOYD-JONES, Hannah. **Selecting data:** what to keep, what to delete? Exeter: University of Exeter, 2013. Disponível em: http://hdl.handle.net/10036/4427. Acesso em: 08 jun. 2020.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. 2020. Disponível em: http://dadosabertos.cnpq.br Acesso em: 08 jun. 2020.

COUTURE, Carol. Archival appraisal: a status report. **Archivaria**, v. 59, Spring 2005, p. 83-107. Disponível em: https://archivaria.ca/index.php/archivaria/article/view/12502/13624. Acesso em: 08 jun. 2020.

EAKER, C. Selection and appraisal of digital research datasets. *In*: KELLAM, L.; THOMPSON, K. (ed.). **Databrarianship**: the academic data librarian in theory and practice. Chicago: American Library Association, 2016. Disponível em: https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1018&context=utk_libpub. Acesso em: 08 jun. 2020.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO. **Plano de Gestão de dados.** 2020. Disponível em: http://www.fapesp.br/gestaodedados. Acesso em: 08 jun. 2020.

HARVEY, R. Instalment on Appraisal and Selection. In: ROSS, S.; DAY, M. (ed.). **DCC Digital Curation Manual**. 2007. Disponível em: https://researchoutput.csu.edu.au/ws/portalfiles/portal/8624308/CSU281574.pdf. Acesso em: 30 mar. 2021

HEUMÜLLER, R. *et al*. Publish or perish, but do not forget your software artifacts. **Empirical Software Engineering**, v. 25, p. 4585–4616, 2020. DOI: https://doi.org/10.1007/s10664-020-09851-6.

HIGGINS, Sarah. Lifecycle of data management. *In*: PRYOR, Graham. **Managing Research Data.** London: Facet Publishing, 2012.

LIN, Dawei *et al*. The TRUST Principles for digital repositories. **Scientific Data**, v. 7, n. 144, 2020.

THE NATIONAL ARCHIVES. **What is appraisal?**. 2013. p. 3. Disponível em: https://www.nationalarchives.gov.uk/documents/information-management/what-is-appraisal.pdf. Acesso em: 30 mar. 2021.

NATURAL ENVIRONMENT RESEARCH COUNCIL. **NERC Data Value Checklist**. [2019]. 3 p. Disponível em: https://nerc.ukri.org/research/sites/data/policy/data-value-checklist/. Acesso em: 08 jun. 2020.

NDSA Agenda Working Group. **2020 NDSA Agenda**. Charlottesville, Center for Open Science, 2020. DOI 10.17605/OSF.IO/BCETD.

PENNOCK, Maureen. **Digital Curation**: a Life-Cycle Approach to Managing and Preserving Usable Digital Information, 2007. Disponível em: http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf. Acesso em: 08 jun. 2020.

PrePARe Project Team. **Selecting what to keep and what to bin (Checklist***).* Cambridge: Cambridge University Library, 2012. 1 p. Disponível em: http://www.dspace.cam.ac.uk/handle/1810/243754. Acesso em: 08 jun. 2020.

SAYÃO, Luis Fernando; SALES, Luana Farias. Curadoria Digital: um novo patamar para preservação de dados digitais de pesquisa**.** **Informação & Sociedade Estudos**, João Pessoa, v. 22, n. 3, p. 179-191, set./dez. 2012.

SAYÃO, Luis Fernando; SALES, Luana Farias. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. **Tendências da Pesquisa Brasileira em Ciência da Informação**, João Pessoa, v. 6, p. 1-26, 2013. Disponível em: http://carpedien.ien.gov.br/bitstream/ien/646/1/DADOS%20DE%20PESQUISA.pdf . Acesso em: 19 fev. 2021.

SIEBRA, Sandra de Albuquerque; et al. Curadoria digital: além da questão da preservação digital. In: **XIV Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB)**, 2013. Disponível em: 200.20.0.78/repositorios/handle/123456789/2478 . Acesso em: 19 fev. 2021.

SIEBRA, Sandra de Albuquerque. O planejamento na curadoria digital. **Informação & Sociedade Estudos**, João Pessoa, v. 30, n. 4, p. 1-22, out./dez. 2020.

UK DATA SERVICE. **Collections Development Selection and Appraisal Criteria**, 2018. Disponível em: https://www.ukdataservice.ac.uk/media/455175/cd234-collections-appraisal.pdf . Acesso em: 20 jun. 2020.

UNIVERSITY OF BRISTOL. Research Data Service. **Research Data Evaluation Guide.** Bristol: 2018. 6 p. Disponível em: http://www.bristol.ac.uk/staff/researchers/data/writing-a-data-management-plan/. Acesso em: 08 jun. 2020.

WHYTE, A.; WILSON, A. **How to appraise & select research data for curation**. Edinburgh, DCC, 2010. Disponível em: http://www.dcc.ac.uk/resources/how-guides/appraise-select-research-data. Acesso em: 08 jun. 2020.

[i] TIBBO, H.R. On the nature and importance of archiving in the digital age. Advances in Computers, 57, p. 1-67, 2003.