Publicando dados de pesquisa: contextualizando as principais etapas e elementos envolvidos no processo

Guilherme Ataíde Dias
Pós-Doutorado pela Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) - Brasil.
Doutor em Ciências da Comunicação /Ciência da Informação pela Universidade de São Paulo (USP) - Brasil.
Professor da Universidade Federal da Paraíba (UFPB) - João Pessoa, PB - Brasil
http://lattes.cnpq.br/9553707435669429
https://orcid.org/0000-0001-6576-0017
E-mail: guilhermeataide@ccsa.ufpb.br

Sandra de Albuquerque Siebra
Doutora em Ciências da Computação pela Universidade Federal de Pernambuco (UFPE) – PE - Brasil.
Professora da Universidade Federal de Pernambuco (UFPE) - Recife, PE - Brasil
http://lattes.cnpq.br/4923627544089379
https://orcid.org/0000-0002-0078-6918
E-mail: sandra.siebra@gmail.com

Rosilene Paiva Marinho de Sousa
Doutora em Ciência da Informação pela Universidade Federal da Paraíba (UFPB) – PB - Brasil.
Professora da Universidade Federal do Oeste da Bahia (UFOB) - Barreiras, BA – Brasil.
http://lattes.cnpq.br/4465533418771961
https://orcid.org/0000-0002-4699-8692
E-mail: adv.rpmarinho@gmail.com

Marckson Roberto Ferreira de Sousa
Doutor em Engenharia Elétrica pela Universidade Federal da Paraíba (UFPB) – PB - Brasil.
Professor da Universidade Federal da Paraíba (UFPB) - João Pessoa, PB – Brasil.
http://lattes.cnpq.br/0221265788966967
https://orcid.org/0000-0003-2001-1631
E-mail: marckson.dci.ufpb@gmail.com

Submetido em: 24/11/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021 .

RESUMO

O uso e reúso de dados de pesquisa são ações importantes e necessárias no contexto atual do fazer científico. Por isso, os conjuntos de dados produzidos durante as iniciativas de pesquisa precisam ser publicados para serem mais facilmente acessados por toda a comunidade científica. Porém, publicar dados não significa a mera disponibilização destes em um repositório, mas compreende uma série de etapas que devem ser consideradas para que eles possam, efetivamente, serem utilizados. Assim, este trabalho tem como objetivo principal discutir o processo de publicação de conjuntos de dados de pesquisa no contexto da ciência. Esta é uma investigação descritiva e qualitativa, que faz uso da pesquisa bibliográfica. Como resultado são apresentados elementos envolvidos no processo de publicação de dados de pesquisa e discussões abrangendo as fases de publicação de conjuntos de dados, relativas ao depósito, descrição, atribuição de identificador e revisão, etapas essas que envolvem o núcleo do processo de publicação.

Palavras-chave: Publicação de dados. Compartilhamento de dados. Artigo de dados. Periódico de dados. Dados científicos.

Publishing research data: contextualizing the main steps and elements involved in the process

ABSTRACT

The use and reuse of research data are important and necessary actions in the current context of scientific practice. For this reason, the datasets produced during the research initiatives need to be published to be more easily accessed by the entire scientific community. However, publishing data does not mean merely making it available in a repository, but comprises a series of steps that must be considered so that they can be effectively used. Thus, this work has as main objective to discuss the process of publishing research datasets in the context of science. This is a descriptive and qualitative investigation, which makes use of bibliographic research. As results, elements involved in the process of publishing research data and discussions are presented, covering the stages of publication of datasets, relating to the deposit, description, identifier assignment and review, steps that involve the core of the publication process.

Keywords*: Data publication. Data sharing. Data paper. Data journal. Scientific data.*

Publicación de datos de investigación: contextualización de los principales pasos y elementos implicados en el proceso

RESUMEN

El uso y reúso de datos de investigación son acciones importantes y necesarias en el contexto actual del hacer científico. Por esta razón, los conjuntos de datos producidos durante las iniciativas de investigación necesitan ser publicados, de tal manera que, toda la comunidad científica pueda acceder fácilmente a ellos. Sin embargo, publicar datos no significa simplemente colocarlos a disposición en un repositorio, sino comprende una serie de fases que deben ser consideradas para que, los datos, puedan utilizarse de manera efectiva. Es así que, este estudio tiene como objetivo principal discutir el proceso de publicación de conjuntos de datos de investigación en el contexto de la ciencia. Se trata de una investigación de tipo descriptiva con enfoque cualitativo, la cual hace uso de investigación bibliográfica. Como resultado, se presentan elementos involucrados en el proceso de publicación de los datos de investigación, y discusiones que abarcan las fases de publicación de los conjuntos de datos, relativas al depósito, la descripción, la asignación del identificador y la revisión, fases que constituyen el núcleo del proceso de publicación.

Palabras clave*: Publicación de datos. Uso compartido de datos. Artículo de datos. Revista de datos. Datos científicos.*

INTRODUÇÃO

Compartilhar dados de pesquisa é uma ação entendida como de fundamental importância pela comunidade científica (BORGMAN, 2015; DRAZEN et al, 2016). Os benefícios associados ao compartilhamento de dados, tais como a redução de custos no processo de investigação; a reprodutibilidade dos experimentos; a possibilidade de comprovação de resultados obtidos; o aumento das colaborações entre pesquisadores e o retorno para a sociedade dos investimentos públicos realizados através de órgãos de fomentos, contribuindo para um melhor impacto socioeconômico das pesquisas públicas, dentre outros, são bastante evidentes. Porém, para os dados de pesquisa terem seu uso potencializado por meio do compartilhamento, carecem de uma publicação eficaz, o que significa não simplesmente disponibilizá-los no servidor de um programa de pós-graduação ou de uma instituição de pesquisa, na página pessoal do próprio pesquisador, ou mesmo em um repositório institucional de uma instituição de pesquisa ou universidade. O processo de publicação de dados de pesquisa pode influenciar no efetivo compartilhamento destes recursos e na ampliação ou não do seu uso e reúso. Neste trabalho adota-se a definição de Publicação de dados (com P maiúsculo) de Callaghan *et al.* (2012), que aponta que a Publicação é um processo formal que deve: fornecer mecanismo para assegurar o crédito ao trabalho do pesquisador (de forma que os dados possam ser formalmente citados); possibilitar que se agregue valor a um conjunto de dados (pois estes precisam estar documentados) e garantir a persistência destes.

Para que a Publicação de dados de pesquisa seja bem sucedida é necessário o envolvimento e conscientização e, às vezes, a capacitação/treinamento da comunidade de pesquisa no processo. E para a motivação desta comunidade, é importante divulgar para os pesquisadores as vantagens que eles podem obter ao adotar a referida prática. Costello (2009) lista uma série de benefícios associados à publicação de dados que podem ser auferidos ao cientista individual responsável pela criação de conjuntos de dados. Estes benefícios, dentre outros, incluem: "Maior taxa de citação"; "Reconhecimento entre os pares ampliado"; "Convites para encontros"; "Convites para publicar"; "Convites para prover consultoria"[1] (COSTELLO, 2009, p. 420), além de se criar oportunidades para parcerias e trabalhos colaborativos.

Modelos de ciclo de vida de dados tem se mostrado úteis para orientar, definir e ilustrar visualmente os processos/etapas pela quais os dados devem passar até a sua Publicação. Estes ciclos englobam atividades/ações a serem realizadas, de acordo com as necessidades específicas de cada pesquisador e os tipos de dados sendo trabalhados, além de papéis e responsabilidades, marcos e outros componentes importantes para a gestão, preservação e disponibilização dos dados (CARLSON, 2014; KOWALCZYK, 2018). Existem diversos modelos de ciclo de vida dos dados, tais como o DCC Curation Lifecycle Model (DCC, 2020), o DataONE Data Life Cycle (DATAONE, 2020), o DDI Combined Life Cycle Model e o Research Data Lifecycle – UK Data Service (UK DATA SERVICE, 2020). Pode ser observado que esses modelos apresentam diferentes níveis de granularidade e detalhes, mas todos são organizados em processos/etapas que possuem semelhanças, sendo as principais: planejamento; criação, coleta ou captura; armazenamento; gestão a longo prazo/preservação; processamento ou análise; Publicação ou compartilhamento; uso e reúso. (CORTI *et al.*, 2014; KOWALCZYK, 2018). Entretanto, estes modelos, se observados sob a ótica dos pesquisadores, podem ser abstratos e apresentar uma certa complexidade para orientar o processo de Publicação dos dados. Assim, um modelo mais recente encontrado na literatura, focado na Publicação de dados, é o apresentado por Kratz e Strasser (2014).

[1] Texto original em inglês:
"Greater citation rate"
"Wider recognition among peers"
"Invitations to meetings"
"Invitations to collaborate"
"Invitations to provide consultancy"

Um modelo que simplifica o processo de Publicação dos conjuntos de dados e, apesar de não englobar todas as etapas relevantes dos modelos de ciclo de vida de dados existentes (tais como os anteriormente citados), tem potencial para deixar mais claro e simples o processo de Publicação de dados para os pesquisadores em geral.

Nesse contexto, o objetivo principal deste trabalho é discutir o processo de Publicação de conjuntos de dados de pesquisa, tomando como referência o modelo apresentado por Kratz e Strasser (2014).

As pesquisas de Dias, Anjos e Araújo (2019) e de RDP BRASIL (2018) revelaram que muitos dos pesquisadores brasileiros de todas as áreas do conhecimento possuem incertezas sobre questões relacionadas aos dados de pesquisa. Assim, espera-se, com esse trabalho, contribuir para o aumento da compreensão da temática por cientistas de todas as áreas do conhecimento, assim como com o despertar para outras questões associadas aos dados de pesquisa.

METODOLOGIA

Do ponto de vista de seu objetivo, esta pesquisa é descritiva. Para Richardson (2017, p. 8), a pesquisa descritiva "[...] procura descrever sistematicamente uma situação, problema, fenômeno ou programa para revelar da estrutura o comportamento de um fenômeno".

Os recursos bibliográficos usados na investigação foram obtidos por meio de consultas diretas e indiretas (consulta às referências dos materiais obtidos) realizadas no Portal de Periódicos da CAPES. Este portal foi selecionado em virtude da possibilidade de acessar de maneira centralizada diversas bases de dados. As consultas foram realizadas nos meses de março e abril do ano de 2020. Na construção da query de busca, em um primeiro momento, não foi colocada nenhuma restrição temporal.

Os descritores selecionados como argumentos para a busca, tanto em língua portuguesa, quanto em língua inglesa, foram: "Publicação de dados"; "Data publication"; "Data Publishing"; "Compartilhamento de dados"; "Data sharing"; "Artigo de dados"; "Data paper"; "Periódico de dados"; "Data Journal"; "Dados científicos"; "Scientific data", verificados no título, resumo e palavras-chaves.

Analisando-se o resultado da busca inicial, foi observado que a quantidade de referências retornadas anteriores ao ano 2000 que versavam sobre a Publicação de dados eram praticamente nulas. Desta forma, refinou-se a *query* de busca estabelecendo como novo critério temporal a busca por referências datadas a partir do ano 2000. Os resumos dos artigos recuperados foram lidos, de forma que foram selecionados para leitura completa e, posteriormente para análise e discussão, os artigos que abordavam temáticas relacionadas à Publicação de dados.

A análise e discussão dos artigos da pesquisa tomou como base o modelo de Publicação de dados desenvolvido por Kratz e Strasser (2014), devido ao seu potencial para tornar o processo mais claro e simples para os pesquisadores. Também pelo fato deste ter sido o único modelo focado em Publicação de dados encontrado nas pesquisas realizadas.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A busca por definições (ou redefinições) em Ciência da Informação acerca de objetos sob seu tratamento é uma das tarefas fundamentais de seus pesquisadores (DIAS; VIEIRA; SILVA, 2013). Assim, ao se considerar a Publicação de dados de pesquisa, há a necessidade de definir, redefinir ou conceituar com maior precisão, o conceito de dados de pesquisa. A Organisation for Economic Co-operation and Development - OECD (Organização para a Cooperação e Desenvolvimento Econômico - OCDE) define dados de pesquisa como

> […] registros factuais (resultados numéricos, registros textuais, imagens e sons) usados como fontes primárias para a pesquisa científica e que são comumente aceitos na comunidade científica como necessários para validar os resultados da pesquisa. Um conjunto de dados de pesquisa constitui uma representação parcial e sistemática do sujeito que está sendo investigado (OECD, 2007, p.13, tradução nossa) [2].

Borgman (2015, tradução nossa, p.28)[3] explica o conceito de dados em um contexto relacionado à pesquisa como sendo "[…] representações de observações, objetos ou outras entidades usadas como evidência de fenômenos para fins acadêmicos ou de pesquisa". Seguindo linha de raciocínio similar, Assante *et al.* (2016) entendem que os dados de pesquisa estão relacionados com uma gama variada de materiais produzidos ao longo de atividades de pesquisa.

Embora não seja objeto desse estudo, pois para isso seria necessário um estudo de terminologia, ressalta-se que, frequentemente, podem ser encontrados nos textos científicos na área da Ciência da Informação o uso das expressões "dados de pesquisa" e "dados científicos" empregadas como sinônimos.

Essas duas expressões apresentam sutilezas que podem levar a compreensões diversas, porém, no contexto deste trabalho, será utilizada, preferencialmente a expressão "dados de pesquisa", na forma como conceituada pela OECD (2007). Sendo utilizada a expressão "dados científicos" apenas por questão de coerência, quando a referência original utilizada assim fizer.

As recomendações para a disponibilização de dados na *Web* (LÓSCIO; BURLE; CALEGARI, 2017; GOFAIR, 2020; OPEN KNOWLEDGE FOUNDATION, 2020) utilizam indistintamente os termos disponibilização, abertura, compartilhamento e publicação de dados, ao se referir ao processo de tornar os dados de pesquisa passíveis de recuperação, disponíveis para acesso, uso e reúso. Porém, nesse trabalho, que tem foco no pesquisador e descreve um processo que requer rigor na sua execução, será utilizado o termo Publicação de dados.

O processo de publicar dados de pesquisa consiste em um conjunto de ações que agrega valor a um conjunto de dados, possibilitando que os mesmos sejam acreditados por uma comunidade específica e que tenham a possibilidade de serem universalmente acessíveis através de ambientes de redes. Este acesso pode ser ao conteúdo integral do conjunto dos dados ou, conforme o esquema de licenciamento utilizado, a um subconjunto dos dados ou, ainda, apenas aos seus metadados descritivos.

Callaghan *et al.* (2013) esclarecem que a Publicação formal de dados agrega serviços aos conjuntos de dados que vão além do simples fato de tê-los disponíveis em um website. Ela inclui verificações de natureza estritamente técnicas, como o tipo do formato dos dados e metadados usados, até considerações de natureza mais científica, tal como verificar se os dados a serem publicados efetivamente possuem significado científico.

Austin *et al.* [4] (2017, p. 82, tradução nossa) definem a Publicação de dados de pesquisa da seguinte forma:

[2] Texto original em inglês: "[…] 'research data' are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.".

[3] Texto original em inglês: "[…] representations of observations, objects, or other entities used as evidence of phenomena for the purpose of research or scholarship.".

[4] Texto original em inglês: "Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way. Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable – all aspects of data publishing that are important for future reuse of data by third party end-users.".

> A publicação de dados de pesquisa é a liberação de dados de pesquisa, metadados associados, documentação acompanhante e código de software (nos casos em que os dados brutos foram processados ou manipulados) para reúso e análise de forma que possam ser descobertos na Web e referenciados de forma única e persistente. A publicação de dados ocorre através de repositórios de dados dedicados e/ou periódicos (de dados) que garantem que os objetos de pesquisa publicados estejam bem documentados, geridos e preservados, arquivados a longo prazo, interoperáveis, citáveis, com qualidade garantida e encontráveis - todos os aspectos importantes da publicação de dados que são importantes para o reúso futuro dos dados por usuários finais terceiros.

Com relação a esta definição de Publicação de dados de pesquisa elaborada por Austin *et al.* (2017), concorda-se com o posicionamento de Dallmeier-Tiessen *et al.* (2017), ao afirmarem que ela é consistente com as etapas necessárias para que um objeto digital de pesquisa esteja em conformidade com os princípios da iniciativa FAIR (Findable, Accessible, Interoperable e Reusable) (WILKINSON *et al.* 2016).

Ou seja, os dados de pesquisa devem ser encontráveis, acessíveis, interoperáveis e reutilizáveis (DIAS, G. A.; ANJOS, R. L.; RODRIGUES, 2019). Desta forma, entende-se *a priori*, que os dados publicados devem estar em sintonia com os princípios preconizados pela iniciativa FAIR (GO FAIR, 2020).

Kratz e Strasser (2014) indicam que a comunidade acadêmica, em sua maioria, concorda que a Publicação de dados é composta por três propriedades fundamentais (Propriedades de publicação), conforme pode ser observado na figura 1, sendo elas: (1) os dados precisam estar disponíveis; (2) os dados devem estar documentados; (3) os dados podem ser citados.

Os autores fazem, ainda, questionamentos acerca de uma quarta propriedade referente a (4) validação dos dados. Indagações acerca de como e em qual medida um conjunto de dados deve ser validado são postas.

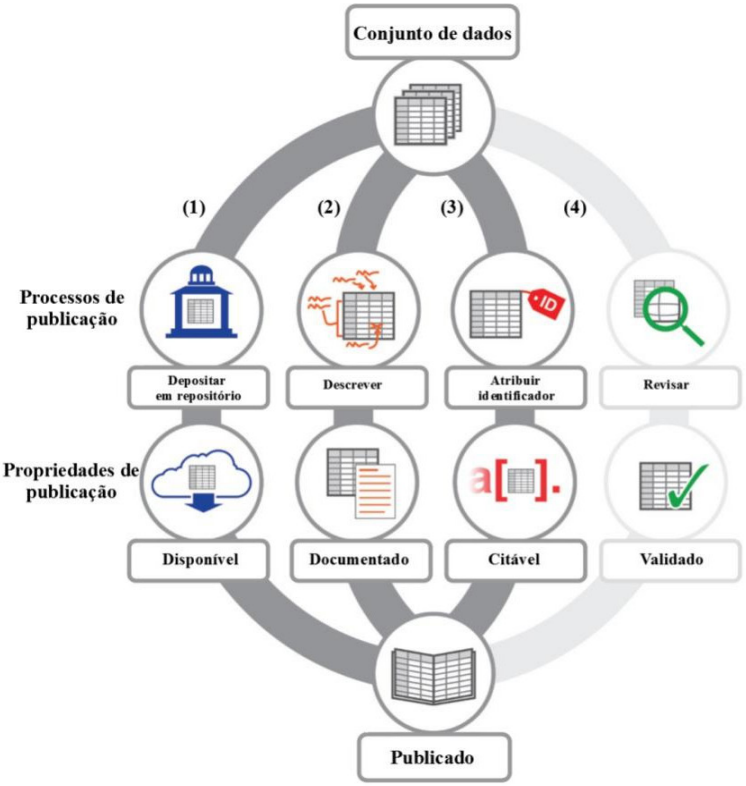
A um primeiro olhar, entende-se que a questão da validação de conjuntos de dados por terceiros no contexto de sua Publicação é uma ação que agrega valor a todo o processo, como será detalhado na subseção "Publicação de Conjuntos de Dados – Revisão".

Os processos de Publicação de dados apresentados por Kratz e Strasser (2014), também presentes na figura 1, estão estruturados em torno de ações que compreendem o (1) depósito de um conjunto de dados em um repositório; (2) sua descrição; (3) atribuição de um identificador; e um eventual (4) processo de revisão, o qual considera-se de fundamental importância. O trabalho de Kratz e Strasser (2014) sobre Publicação de dados pode ser compreendido como um modelo que pode ser ampliado e adaptado a casos concretos.

Alerta-se que, embora os Ciclos de Vida de Dados (CVDs) não sejam o objetivo desse trabalho, as ações que compreendem a Publicação de dados podem ser consideradas como subconjuntos das etapas existentes nos CVDs (ex: Modelo de Ciclo de Vida do DCC, DataONE, DDI, etc). A pesquisa de Araújo *et al.* (2019) apresenta uma compilação das etapas de três CVDs e pode ser utilizada para fazer um paralelo com as ações do modelo apresentado por Kratz e Strasser (2014).

Na sequência são apresentadas as etapas relativas aos processos de Publicação, com base no trabalho de Kratz e Strasser (2014), apresentando-se uma discussão sobre cada uma destas e alguns elementos de seu modelo.

Figura 1 – Publicação de conjuntos de dados



Fonte: Traduzido e adaptado de Kratz e Strasser (2014, p. 3).

PUBLICAÇÃO DE CONJUNTOS DE DADOS – DEPÓSITO

O depósito de um conjunto de dados constitui-se no processo de armazenar esses recursos em um repositório, preferencialmente em um repositório especificamente projetado para o depósito de dados de pesquisa. A disponibilização do conjunto de dados para a comunidade será efetivada após esta ação. Contudo, o depósito dos conjuntos de dados em um repositório de dados não implica que os mesmos estejam automaticamente disponíveis para qualquer pessoa. Estes dados podem estar protegidos por questões de cunho legal, como no caso de conteúdos acerca da saúde de pessoas, ou mesmo estar embargados por um período de tempo pelos seus criadores (KRATZ; STRASSER, 2014). Porém, mesmo no caso em que os conjuntos de dados não estão disponíveis, espera-se que um mínimo de informações estejam disponibilizadas acerca do seu conteúdo por meio do uso de metadados descritivos.

Os repositórios de dados científicos podem ser entendidos como instrumentos que proveem suporte para a Publicação de dados, disponibilizando recursos para todos os atores envolvidos no processo (ASSANTE *et al.*, 2016). Com relação a esses repositórios, Marcial e Hemminger (2010) esclarecem que os mesmos são conceitualmente similares aos repositórios institucionais, embora possuam naturezas muito diferentes no que tange a especificidades de domínio e em alguma medida no que diz respeito à utilidade.

No contexto da Publicação de dados, Curty e Aventurier (2017, p. 6, grifo nosso) trazem o seguinte comentário: "Os repositórios de dados são parte essencial do ecossistema da publicação de dados, e constituem por si só como uma abordagem de data publishing, uma vez que tornam públicas coleções de dados acompanhadas por recursos que otimizem seu potencial de reúso.".

Exemplos de produtos de software que podem ser utilizados para a construção de repositórios de dados incluem o Dataverse[5] e o DSpace[6]. Como implementações efetivas de serviços para o depósito de dados de pesquisa podem ser indicados o Dryad[7] , o Figshare[8] e o Zenodo[9] (ASSANTE *et al.*, 2016). No Brasil, ainda são poucas as instituições de ensino e pesquisa que disponibilizam serviços para o depósito de dados de pesquisa, foi detectado em uma investigação preliminar, ainda em andamento, que nem 20% das universidades federais possuem repositórios de dados.

Ressalta-se que na escolha do software, como afirma Torino, Roa-Martínez e Vidotti (2020), deve ser dada prioridade aos repositórios confiáveis, que possibilitem a preservação digital, a fim de garantir o acesso, uso e reúso dos dados a longo prazo. E, também, conforme o pensamento de Lóscio, Burle e Calegari (2017), que o repositório permita a utilização de padrões de metadados internacionalmente aceitos, a fim de possibilitar interoperabilidade e que os dados possam ser facilmente encontrados.

PUBLICAÇÃO DE CONJUNTOS DE DADOS – DESCRIÇÃO

O processo de descrição consiste em agregar informações adicionais aos conjuntos de dados, de modo a ampliar as possibilidades de interpretações sintáticas e semânticas destes objetos, tanto por seres humanos, quanto por sistemas automatizados, sendo este conjunto de informações que representa a documentação dos conjuntos de dados. Pode-se dizer que essas informações dão o contexto que proporcionará um maior e melhor entendimento sobre o conjunto de dados.

Existem vários tipos de insumos que podem contribuir na composição da documentação de um conjunto de dados. Neste sentido, Kratz e Strasser (2014) elencam os seguintes recursos: os metadados, o artigo cientifico tradicional e o artigo de dados. A estes recursos,sugere-se a inclusão, ainda, de ontologias e os Planos de Gestão de Dados (PGDs).

Ainda, no que diz respeito à elaboração dos metadados, indica-se que estes devem estar em sintonia com os princípios da iniciativa FAIR, contribuindo assim, para que sistemas computacionais possam achar, acessar, interoperar e reusar conjuntos de dados com um menor grau de intervenção humana (GOFAIR, 2020).

Um recurso que pode integrar a documentação de conjuntos de dados e que se entende ser essencial no processo de sua descrição são os artigos de dados. Chavan e Penev[10] (2011, p. 3, tradução nossa) explicam o que vem a ser um artigo de dados:

> Um artigo de dados é uma publicação de um periódico cujo objetivo é descrever dados, ao invés de relatar uma investigação. Como tal, contém fatos sobre dados, não hipóteses e argumentos em apoio a essas hipóteses baseadas em dados, conforme encontrado em um artigo de pesquisa convencional.

O que é complementado pela definição dada por Torino, Roa-Martínez e Vidotti (2020, p. 188), que afirma que um artigo de dados é "um objeto científico que descreve minuciosamente todos os elementos necessários para a compreensão do conjunto de dados, incluindo a justificativa e os métodos de coleta". Este objeto pode ser publicado em um periódico convencional ou em um periódico de dados (data jornal).

[5] Disponível para download a partir de https://dataverse.org/
[6] Disponível para download a partir de https://duraspace.org/dspace/
[7] Disponível a partir de https://datadryad.org/
[8] Disponível a partir de https://figshare.com/
[9] Disponível a partir de https://zenodo.org/

[10] "A data paper is a journal publication whose primary purpose is to describe data, rather than to report a research investigation. As such, it contains facts about data, not hypotheses and arguments in support of those hypotheses based on data, as found in a conventional research article.".

Os artigos de dados são importantes pelo fato de servirem como instrumentos que possibilitam um aumento na visibilidade dos dados, contribuindo, portanto, para o uso e reúso, e para que estes possam ser utilizados inúmeras vezes, inclusive para objetivos distintos, pois um cientista pode compreender mais facilmente o que um determinado conjunto de dados representa pela leitura de um artigo de dados associado, uma vez que ele conterá informações que vão desde os formatos de dados e sua localização, até a modalidade de política autoral adotada pelos detentores (SANT'ANA, 2019).

Assim, os artigos de dados além de contribuir para que dados sejam mais visíveis e passíveis de uso, reúso e compartilhamento, servem para divulgação da modalidade de política de direito autoral adotada pelo mesmo (SOUSA; DIAS; SOUSA, 2020). De modo análogo ao artigo científico tradicional, o artigo de dados pode trazer prestígio e reconhecimento da comunidade científica para seus autores. Para um maior aprofundamento nas questões que permeiam os artigos de dados, indica-se a pesquisa de Roa-Martinez *et al.* (2017), que propõe a estrutura comum de um artigo de dados baseada em conjuntos de metadados.

PUBLICAÇÃO DE CONJUNTOS DE DADOS – ATRIBUIÇÃO DE IDENTIFICADOR

Equivalente ao que acontece com artigos científicos, o uso de identificadores também pode ser usado para reconhecer de forma única um conjunto de dados. A persistência de um identificador é uma característica fundamental, pois permite que os conjuntos de dados sejam acessados de forma inequívoca, possibilitando a recuperação destes a partir de um endereço padrão associado ao identificador. É necessário mencionar que sempre que a localização de um determinado conjunto de dados for alterada, esta mudança deve ser refletida no identificador que referencia o respectivo recurso. Neste diapasão, Lawrence *et al.* (2011) indicam que o uso de um identificador é uma solução possível para abordar o problema de permanência no meio eletrônico.

A partir do momento em que o pesquisador dispõe de um identificador para um conjunto de dados, torna-se possível a sua citação inequívoca. Citar um conjunto de dados no artigo científico que resultou do seu uso possibilita que terceiros acessem este conjunto de dados e tenham a oportunidade de reutilizá-los. Ressalta-se que a citação de um conjunto de dados confere reconhecimento aos seus respectivos criadores.

Um objeto identificador pode ser utilizado como mecanismo no controle dos direitos autorais de produção intelectual em meio eletrônico, por permitir o uso de um único código de identificação para objetos usados em redes digitais. A Lei de Direitos Autorais (BRASIL, 1998) não contempla em sua totalidade as diversidades de criação do intelecto humano que surgem no ambiente atual, a exemplo dos bens intelectuais advindos das tecnologias de informação e comunicação (SOUSA; DIAS, 2012). Diante disso, um identificador pode ser visto como uma alternativa que pode contribuir na proteção do direito autoral no que diz respeito ao conteúdo, resguardando desse modo todas as características e atributos inerentes aos direitos morais e patrimoniais previstos em lei.

O Digital Object Identifier (DOI) é um exemplo de identificador que pode ser usado no contexto dos conjuntos de dados. Serviços de repositórios de dados como Dryad, o Zenodo e outros atribuem um DOI para cada conjunto de dados neles depositados. A figura 2 ilustra um exemplo de uma referência para um conjunto de dados depositado no Zenodo que dispõe sobre o número de casos de COVID-19 na Irlanda. O fragmento de texto destacado na figura 2 corresponde (10.5281/zenodo.3754983) ao DOI associado ao conjunto de dados. A obtenção do DOI pode ser realizada através de uma agência de registro. Duas das agências mais conhecidas são a Crossref [11] e a DataCite[12] . Ambas as agências proveem identificadores, contudo elas possuem objetivos distintos.

[11] Disponível a partir de: https://www.crossref.org/

[12] Disponível a partir de: https://datacite.org/

Figura 2 − Exemplo de citação para conjunto de dados

> Moriarty, Frank. (2020). Number of cases of coronavirus disease (COVID-19) in Ireland (Version v2.0) [Data set]. Zenodo. http://doi.org/**10.5281/zenodo.3754983**

Fonte: Moriarty (2020, grifo nosso).

A Crossref tem foco em atribuir identificadores para artigos de periódicos, artigos de conferências, livros e outros materiais, enquanto que o DataCite, por estar envolvido em uma ecologia mais voltada para a questão dos dados de pesquisa, está mais centrado neste tipo de objeto (CROSSREF, 2020). Assim, torna-se relevante a existência de um identificador único no processo de Publicação de conjuntos de dados, para resguardar a atribuição de autoria.

PUBLICAÇÃO DE CONJUNTOS DE DADOS – REVISÃO

A revisão de conjuntos de dados para a sua validação e posterior Publicação é uma etapa importante e que pode conferir confiança no uso destes recursos por terceiros. O processo de validação de conjunto de dados poderia, em alguma medida, mimetizar o processo de revisão pelos pares (*peer-review*) utilizado na revisão de artigos científicos (ASSANTE, 2016). Não obstante entender esta revisão como uma etapa necessária, questionamentos e dúvidas emergem com relação de como seria efetivamente este processo de validação (CALLAGHAN, 2013; KRATZ; STRASSER, 2014; ASSANTE, 2016).

É necessário refletir em que medida um eventual processo de revisão por pares de conjuntos de dados seria diferente do processo de revisão de artigos científicos, de modo que várias considerações podem emergir. Considera-se perfeitamente factível que um determinado parecerista, ao revisar um artigo que verse sobre sua área de domínio conhecimento, faça-o sem maiores dificuldades. Porém, no caso de ser atribuído para um parecerista um conjunto de dados, mesmo que ele entenda dos fatos representados pelos dados, pode ser que não entenda dos aspectos técnicos associados aos mesmos. Estes aspectos incluem: tipos de metadados usados, formatos de dados e produtos de software relacionados ao processo de captura e tratamento dos dados. Salienta-se ainda a questão de sobrecarga do trabalho a que os pesquisadores estão submetidos.

O tempo disponível para a revisão de artigos científicos é consideravelmente reduzido e compete com várias outras atividades de pesquisa com a qual estão envolvidos, assim, torna-se custoso ter de acoplar ainda. Desta forma, como seria possível equalizar a agenda de trabalho para ainda alocar tempo direcionado para a revisão de conjuntos de dados?

Callaghan (2013) comenta que, no ano de 2010, o periódico intitulado Journal of Neuroscience suspendeu a adição de qualquer arquivo suplementar ao artigo principal, visto a percepção de que os revisores estavam "imersos" pelo volume de dados e outros arquivos para avaliar, além do texto do artigo principal. Ainda com relação ao processo de revisão, Callaghan (2013) fez uma série de questionamentos. O autor refletiu sobre o que especificamente o parecerista deveria revisar e questionou se o que deve ser avaliado são os dados brutos, os metadados, o artigo de dados ou todos os elementos mencionados. O autor complementa indagando sobre que tarefa efetivamente deveria ser solicitada para o revisor executar. Outro ponto a considerar é que, ao contrário de um artigo científico, que uma vez publicado através de um periódico, normalmente não está sujeito a modificações, os conjuntos de dados podem passar por múltiplas versões ao longo do seu ciclo de vida. Desta forma, é possível levantar algumas questões adicionais sobre o processo de revisão de dados: como seria a revisão de um conjunto de dados que possui múltiplas versões e quem seria o parecerista em cada uma das versões dos artigos?

Sugestões para alguns dos questionamentos aqui postos são discutidos nos trabalhos de Klump *et al.* (2006), Lawrence *et al.* (2011) e Parsons e Fox (2013).

Os referidos autores apresentam ideias relacionadas, por exemplo, considerar a data da versão dos dados, onde o processo de revisão pode inclusive considerar aspectos da qualidade dos dados mantidos, bem como de seus metadados associados, devendo-se cada versão ser utilizada com bastante cautela.

Os revisores poderiam também incluir comentários sobre os dados que resultariam em possibilidade de mudanças em ambos, o que pode exigir que tanto os dados, quanto os metadados, possuam sempre novas versões correspondentes, que seriam publicadas em vez dos dados e metadados originais. Porém, esse processo não apresenta simplicidade, pois conduz a uma necessidade de se ter que solicitar aos autores que respondam aos revisores por mudanças ocorridas tanto nos dados, quanto nos metadados, podendo até mesmo requerer que os ajustes realizados por um curador também possam ser atualizados em reflexo ao processo de revisão. Todo esse processo pode ter necessidades específicas ao se considerar a devida proteção nos dados. Além, caso exista necessidade, da atribuição de direitos autorais, fato este que pode considerar diferenças culturais ou diferentes paradigmas de produção, a exemplo não só da qualidade dos dados, mas também dos aspectos espacial e temporal. Adicionalmente, considere-se incluso no processo o aumento de trabalho do parecerista/revisor para acompanhar as mudanças e ajustes.

A situação relativa à Publicação de dados pode apresentar ainda barreiras informacionais quando do compartilhamento dos dados, necessitando de procedimentos bem definidos para evitar inconsistências e o consequente desencontro do respectivo conteúdo, caso não se controle a persistência e confiabilidade dos dados e metadados subjacentes.

CONSIDERAÇÕES FINAIS

A partir das análises realizadas fica evidente que o processo de Publicação de dados não é uma atividade simples, mas, ao mesmo tempo, configura-se como primordial para o uso e o reúso de dados de pesquisa.

O domínio das atividades associadas com cada etapa da Publicação de dados (depósito, descrição, atribuição de identificador e revisão) pode transcender os saberes de uma única área do conhecimento. Desta forma, torna-se patente a necessidade de ter-se uma equipe composta por especialistas de diversas áreas, para que se possa efetivamente abordar a questão da Publicação de dados de pesquisa. Pois é improvável que um pesquisador domine, ao mesmo tempo, questões que envolvam repositórios de dados, uso de metadados, atribuição de identificadores, revisão de conjuntos de dados e mais uma série de questões tecnológicas subjacentes ao processo. Dentre os profissionais que poderiam contribuir com esta iniciativa podemos indicar, em um primeiro momento, o pessoal vinculado com as tecnologias da informação e comunicação, bibliotecários e arquivistas. Esta lista não é exaustiva, outros profissionais podem ser incluídos no processo, a partir das demandas específicas de cada caso concreto. Um caminho para a efetivação do processo de Publicação de dados para a comunidade de pesquisa passa pelo apoio das instituições a que estes pesquisadores estão vinculados. Isso demanda o estabelecimento de políticas relacionadas com a questão de dados no âmbito institucional, passando também pelo apoio (financeiro, de infraestrutura, etc.) das organizações governamentais relacionados ao desenvolvimento da ciência e tecnologia.

O assunto tratado nessa pesquisa tem potencial para gerar diversas outras pesquisas. Uma possibilidade de investigação futura seria avaliar o grau de maturidade das organizações de pesquisa no que diz respeito às etapas associadas com o processo de Publicação de dados.

REFERÊNCIAS

ARAÚJO, D. G. *et al.* Contribuições para a gestão de dados científicos: análise comparativa entre modelos de ciclo de vida dos dados. *Liinc em Revista,* Rio de Janeiro, v. 15, n. 2, p. 32-51, nov. 2019. Disponível em: https://doi.org/10.18617/liinc.v15i2.4686. Acesso em: 21 abr. 2020.

ASSANTE, M. *et al.* Are scientific data repositories coping with research data publishing?. *Data Science Journal,* [*s. l.*], v. 15, n. 6, p. 79–83, 2016. Disponível em: http://dx.doi.org/10.5334/dsj-2016-006. Acesso em: 12 abr. 2020.

AUSTIN, C. *et al.* Key components of data publishing: using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries,* [*s. l.*], v. 18, n. 2, p. 77–92, 2017. Disponível em: http://search-ebscohost-com.ez15.periodicos.capes.gov.br/login.aspx?direct=true&db=lih&AN=122919195&lang=pt-br&site=ehost-live. Acesso em: 21 abr. 2020.

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

BRASIL. Lei nº 9.610, de 19 de fevereiro de 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. *Diário Oficial da União*, Brasília, DF, 20 fev. 1998. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L9610.htm. Acesso em: 27 abr. 2020.

CALLAGHAN, S. *et al.* Making data a first class scientific output: data citation and publication by NERC's Environmental Data Centres. *International Journal of Digital Curation,* [*s. l.*], v. 7, n. 1, p. 107–113, 2012. Disponível em: https://doi.org/10.2218/ijdc.v7i1.218. Acesso em: 12 abr. 2020.

CALLAGHAN, S. *et al.* Processes and procedures for data publication: a case study in the geosciences. *International Journal of Digital Curation,* [*s. l.*], v. 8, n. 1, p. 193-203, 2013. Disponível em: https://doi.org/10.2218/ijdc.v8i1.253. Acesso em: 12 abr. 2020.

CARLSON, J. The use of life cycle models in developing and supporting data services. In: RAY, J. M. (org.). *Research Data Management*: practical strategies for information professionals. West Lafayette: Purdue University Press, 2014. p. 63–86.

CHAVAN, V.; PENEV, L. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics,* [*s. l.*], v. 12, n, 15, 2011. Disponível em: https://doi.org/10.1186/1471-2105-12-S15-S2. Acesso em: 12 abr. 2020.

CORTI, L. *et al*. *Managing and sharing research data*: a guide to good practice. Los Angeles: Sage, 2014. 234 p.

COSTELLO, M. J. Motivating Online Publication of Data. *BioScience,* [*s. l.*], v. 59, n. 5, p. 418–427, 2009. Disponível em: http://dx.doi.org/10.1525/bio.2009.59.5.9. Acesso em: 10 abr. 2020.

CROSSREF. *The basics.* 2020. Disponível em: https://www.crossref.org/community/datacite/. Acesso em: 10 abr. 2020.

CURTY, R.; AVENTURIER, P. O paradigma da publicação de dados e suas diferentes abordagens. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 18., 2017, Marília, *Anais* [...]. Marília, SP. Disponível em: https://hal.archives-ouvertes.fr/hal-01637987. Acesso em: 15 abr. 2020.

DALLMEIER-TIESSEN, S. *et al.* Connecting data publication to the research workflow: a Preliminary analysis. *International Journal of Digital Curation,* [*s. l.*], v. 12, n. 1, p. 88-105, 2017. Disponível em: https://doi.org/10.2218/ijdc.v12i1.533. Acesso em: 13 abr. 2020.

DATA DOCUMENTATION INITIATIVE (DDI). *Why Use DDI*. [*S. l.*], 2020. Disponível em: http://www.ddialliance.org/training/why-use-ddi. Acesso em: 13 abr. 2020.

DATA OBSERVATION NETWORK FOR EARTH (DataOne). *Data Life Cycle.* [*S.l.*], 2020. Disponível em: https://www.dataone.org/data-life-cycle. Acesso em: 13 abr. 2020.

DIAS, G. A.; ANJOS, R. L.; RODRIGUES, A. A. Os princípios FAIR: viabilizando o reúso de dados científicos. *In*: DIAS, A. D.; OLIVEIRA, B. M. J. F (Orgs.). *Dados científicos:* perspectivas e desafios. João Pessoa: Editora UFPB, 2019. p. 177-187.

DIAS, G. A.; ANJOS, R. L.; ARAUJO, D. G. A gestão dos dados de pesquisa no âmbito da comunidade dos pesquisadores vinculados aos programas de pós-graduação brasileiros na área da Ciência da Informação: desvendando as práticas e percepções associadas ao uso e reúso de dados. *Liinc em Revista*, Rio de Janeiro, v. 15, n. 2, p. 5-31, 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4683. Acesso em: 21 abr. 2020.

DIAS, G. A.; VIEIRA, A. A. N.; SILVA, A. L. A. Em busca de uma definição para o livro eletrônico: o conteúdo informacional e o suporte físico como elementos indissociáveis. In *Encontro Nacional de Pesquisa em Ciência da Informação,* 2013, Florianópolis. Disponível em: http://eprints.rclis.org/20904/. Acesso em: 10 abr. 2020.

DIGITAL CURATION CENTRE (DCC). *DCC Curation Lifecycle Model.* 2018. Disponível em: https://www.dcc.ac.uk/guidance/curation-lifecycle-model .Acesso em: 12 abr. 2020.

DRAZEN, J. M. *et al.* The importance and the complexities of data sharing. *The New England Journal of Medicine,* [*s. l.*], v. 375, n. 12, p. 1182-1183, 2016. Disponível em: https://www-nejm-org.ez15.periodicos.capes.gov.br/doi/10.1056/NEJMe1611027. Acesso em: 28 abr. 2020.

GOFAIR. FAIR *principles*. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 16 abr. 2020.

KLUMP, J. *et al.* Data publication in the open access initiative. *Data Science Journal,* [*s. l.*], v. 5, p. 79–83, 2006. Disponível em: http://doi.org/10.2481/dsj.5.79. Acesso em: 10 abr. 2020.

KOWALCZYK, S. T. *Digital Curation for Libraries and Archives*. Santa Barbara, California: Libraries Unlimited, 2018.

KRATZ, J.; STRASSER, C. Data publication consensus and controversies. *F1000Research*, [s. l.], v. 3, n. 94, p. 1-21, 2014. Disponível em: https://doi.org/10.12688/f1000research.3979.3. Acesso em: 9 abr. 2020.

LAWRENCE, B. *et al.* Citation and peer review of data: moving towards formal data publication. *International Journal of Digital Curation*, [s. l.], v. 6, n. 2, p. 4-37, 2011. Disponível em: http://www.ijdc.net/article/view/181/265. Acesso em: 10 abr. 2020.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. (ed.). *Data on the web best practices*. 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 14 jan. 2020.

MORIARTY, F. Number of cases of coronavirus disease (COVID-19) in Ireland. *Zenodo*, [*s. l.*], v. 2, 2020. Disponível em: http://doi.org/10.5281/zenodo.3754983. Acesso em: 10 abr. 2020.

MARCIAL, L. H.; HEMMINGER, B.M. Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology,* [*s. l.*], v. 61, n. 10, p. 2029-2048, 2010. Disponível em: https://doi-org.ez15.periodicos.capes.gov.br/10.1002/asi.21339. Acesso em: 14 abr. 2020.

OECD. Organisation for Economic Co-Operation and Development. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. OECD Publishing, Paris, 2007. Disponível em: https://www.oecd.org/sti/inno/38500813.pdf. Acesso em: 10 abr. 2020.

PARSONS, M. A.; FOX, P. A. Is data publication the right metaphor?. *Data Science Journal,* [*s. l.*], v. 12, p. WDS32–WDS46, 2013. Disponível em: https://doi.org/10.2481/dsj.WDS-042. Acesso em: 13 abr. 2020.

ROA-MARTINEZ, S. M. *et al.* Estructura propuesta del artículo de datos como publicación científica. *Revista Española de Documentación Científica,* [*s. l.*], v. 40, n. 1, p.1-12, 2017. Disponível em: http://dx.doi.org/10.3989/redc.2017.1.1375. Acesso em: 14 abr. 2020.

RDP BRASIL. Práticas e Percepções dos Pesquisadores Brasileiros. *Repositórios Piloto da Rede Nacional de Ensino e Pesquisa*, v. 2, 2019. Disponível em: https://hdl.handle.net/20.500.12401/4. Acesso em: 21abr. 2020.

RICHARDSON, R. J. *Pesquisa social*: métodos e técnicas. São Paulo: Atlas, 2017.

SANT'ANA, R. C. G. Campo informacional resultante da interação de ciclos de vida dos dados. In: DIAS, G. A.; OLIVEIRA, B. M. J. F. (Orgs.) *Dados Científicos:* perspectivas e desafios. 1. ed., UFPB: João Pessoa, 2019. p. 33-52.

SOUSA, R. P. M.; DIAS, G. A.; SOUSA, M. R. F. Análise sobre dados abertos e regulação autoral no contexto da editoria científica. In: SHINTAKU, M. *et al.* (Orgs). *Tópicos sobre dados abertos para editores científicos*. Botucatu, SP: ABEC, 2020. 240 p. DOI: 10.21452/978-85-93910-04-3. p. 119-135. Disponível em: https://www.abecbrasil.org.br/arquivos/Topicos_dados_abertos_editores_cientificos.pdf. Acesso em: 14 abr. 2020.

SOUSA, R. P. M.; DIAS, G. A. Digital Object Identifier: uma breve reflexão sobre sua contribuição para proteção do direito autoral de obras literárias no meio digital. *In*: ALBUQUERQUE, M. E. B. C. *et al.* (Orgs.). *Representação da Informação*: um universo multifacetado. João Pessoa: Editora da UFPB, 2012. p. 141-156.

TORINO, E.; ROA-MARTÍNEZ, S. M.; VIDOTTI, S. A. B. G. Dados de pesquisa: disponibilização ou publicação?. In: SHINTAKU, M.; SALES, L. F; COSTA, M. (org).T*ópicos sobre dados abertos para editores científicos.* Botucatu, SP: ABEC, 2020. p. 183-201. DOI: 10.21452/ 978-85-93910-04-3.cap15.

UK DATA SERVICE. *Research data lifecycle.* [*S. l.*], [2020]. Disponível em: https://www.ukdataservice.ac.uk/manage-data/lifecycle. Acesso em: 13 abr. 2020.

WILKINSON, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data.* [*s. l.*], 2016. Disponível em: https://doi.org/10.1038/sdata.2016.18. Acesso em: 13 abr. 2020.