XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação

O PARADIGMA DA PUBLICAÇÃO DE DADOS E SUAS DIFERENTES ABORDAGENS

Renata Gonçalves Curty (Universidade Estadual de Londrina - UEL)

Pascal Aventurier (Institut National de la Recherche Agronomique - INRA)

THE DATA PUBLISHING PARADIGM AND ITS DIFFERENT APPROACHES

Modalidade da Apresentação: Comunicação Oral

Resumo: A ciência aberta vem ditando uma nova dinâmica ao ecossistema da comunicação científica, por meio da reconfiguração e da sofisticação dos métodos para o compartilhamento e o reuso de resultados de pesquisa. Os formatos de publicação tradicionalmente aceitos pela comunidade científica para a disseminação de avanços em pesquisa têm se mostrado inadequados para as novas formas de se comunicar ciência. Dados de pesquisa têm assumido um maior protagonismo, e passaram a ser valorizados como ativos de pesquisa autônomos, de alto valor intrínseco, e publicáveis. Considerando esse cenário, o presente trabalho de revisão tem por objetivo apresentar o atual contexto da publicação de dados e seus impactos na produção e na comunicação científica, tomando como base exemplos e iniciativas vigentes. São apresentadas e discutidas três abordagens para publicação de dados, bem como suas principais vantagens e limitações: os repositórios de dados, as publicações ampliadas, e os artigos de dados, sendo que estes últimos podem ser publicados em periódicos híbridos ou periódicos dedicados à publicação de dados, e ainda elaborados manualmente ou a partir de ferramentas automáticas disponibilizadas por repositórios de dados. Por fim, são apresentados alguns dos desafios que ainda permeiam as discussões sobre publicações de dados, e sinalizadas possibilidades para pesquisas futuras.

Palavras-chave: Artigos de Dados; Ciência Aberta; Publicações Ampliadas; Periódicos de Dados; Reuso de Dados.

Abstract: Open science has given a new impetus to the ecosystem of scientific communication, through the reconfiguration and sophistication of methods for research outputs' sharing and reuse. Traditionally scientific publication formats have become obsolete and inappropriate to the new ways of communicating science. Research data have assumed a greater role, and become valued as autonomous research assets, of high intrinsic value, and publishable. This review paper aims to present the current context of data publication and its impacts on scientific production and communication, based on examples and current initiatives. Three approaches to data publication, as well as their main advantages and limitations, are presented and discussed: data repositories, extended publications, and

data articles, these last which can be published in hybrid journals or journals dedicated exclusively to data publication, manually produced or generated by automatic tools provided by data repositories. Finally, we present some of the challenges that still permeate discussions about data publications, and indicate possibilities for future research.

Keywords: Data Papers; Open Science; Enhanced Publications; Data Journals; Data Reuse.

1 CIÊNCIA ABERTA E O ADVENTO DO PARADIGMA *DATA PUBLISHING*

Na atual conjuntura da comunicação científica, experimentamos uma maior valorização quanto à disponibilidade de dados científicos, de modo a primar pela transparência e reprodutibilidade em pesquisas; aspectos estes que têm sido reavivados a partir de tecnologias e recursos para compartilhamento de dados promovidos por meio das iniciativas da chamada ciência aberta ( ). Os princípios da ciência aberta estão enraizados na premissa de que os resultados da investigação com financiamento público são bens públicos tangíveis que devem estar disponíveis gratuitamente para reuso[1] (DALRYMPLE, 2003).

A ciência aberta ganhou impulso com o avanço de ambientes digitais para a aquisição, arquivamento, manipulação e transmissão de grandes volumes de dados (FRY; SCHROEDER; DEN BESTEN, 2009). Ferramentas computacionais avançadas para o compartilhamento e distribuição de dados estão pavimentando o caminho para uma melhor reprodutibilidade nas investigações científicas.

Um argumento central da ciência aberta é que os dados podem ter várias aplicações e usos, além daqueles que foram inicialmente previstos pelos investigadores originais/primários (EVANS, REIMER, 2009; UHLIR; SCHRÖDER, 2007; VISION, 2010). O compartilhamento de dados é baseado na suposição de que os dados podem ser úteis para os outros - tanto dentro de um mesmo domínio disciplinar como de forma interdisciplinar - e, portanto, ampliam as chances de novos resultados e conhecimentos científicos decorrentes dos dados disponíveis (WALLIS; ROLANDO; BORGMAN, 2013; TENOPIR , 2011).

O movimento em prol da ciência aberta também endossa a chamada equidade ou integridade dos dados, que engloba quatro premissas: 1) Encontrabilidade: dados devem ser possíveis de serem encontrados; 2) Acessibilidade: dados devem ser acessíveis; 3)

[1] Toda nova aplicação de dados por meio de re-análise e replicação, ou como a combinação de diferentes conjuntos de dados por meio de integração ou meta-análise, a partir de novas perguntas de pesquisa, novos métodos de análise, com propósitos similares ou distintos daqueles empreendidos no estudo original, com ou sem a participação do reutilizador (CURTY, 2015).

Interoperabilidade: dados devem ser interoperáveis, e 4) Reuso: Dados devem ser reutilizáveis (WILKINSON , 2016). Tais premissas orientam o princípio FAIR de gestão de dados de pesquisa adotado por diferentes agências de fomento e institutos de pesquisa como:

(NIH), , ,

, , para citar apenas alguns.

Os investimentos para o compartilhamento de dados e para a realização da ciência aberta justificam-se, portanto, pelo potencial de não apenas ampliar a encontrabilidade, a acessibilidade e a condição de manipulação de dados, mas a efetiva reutilização destes ativos. Neste sentido, a sustentabilidade do ciclo da ciência aberta depende da busca por formas eficientes de maximizar o reuso de dados científicos, ao invés de meramente estocá-los como volumes ociosos em repositórios.

A intensificação da disponibilidade de dados reformula o da ciência, criando novas formas para a disseminação dos dados resultantes de pesquisa. Déficits estruturais das publicações científicas tradicionais que comprometem a concretização da ciência aberta (KLUMP , 2006) vêm sendo repensados no atual paradigma da publicação científica, redefinindo os veículos para disseminação científica, com foco nos dados, conhecido por . De modo geral, o termo "publicação de dados" significa que dados estão disponíveis publicamente e são passíveis de citação por qualquer interessado (KRATZ; STRASSER, 2014).

Austin (2016) descrevem o como a publicação e a difusão de dados científicos acompanhados por metadados associados, documentação e código de (em casos de dados brutos processados ou manipulados) para possível reutilização e análise, de modo que possam ser descobertos via web e referenciados de modo único e persistente. Essa abordagem visa disponibilizar dados em forma de publicação de modo a ampliar as condições de reuso dos dados, bem como de atribuição aos seus autores/produtores. Os dados podem ser publicados como materiais suplementares em publicações ampliadas, em artigos de dados ( ) publicados em periódicos dedicados exclusivamente aos artigos de dados ( ), ou mesmo diretamente por meio de ferramentas e recursos disponíveis em repositórios de dados (AUSTIN ., 2016).

Este trabalho de revisão tem por objetivo contextualizar o cenário da publicação de dados ( ) e seus impactos na produção e comunicação científica, tomando como base exemplos e iniciativas vigentes. Discussões recentes de agências governamentais nacionais sobre a gestão de dados científicos sinalizam para uma iminente movimentação em

torno de políticas e iniciativas que valorizem a publicação de dados e seu potencial de reuso. A partir da caracterização desse novo paradigma da publicação científica, e da discussão das diferentes abordagens para a publicação de dados, espera-se contribuir não apenas para um melhor esclarecimento conceitual acerca do tema, mas também com informações que dêem suporte a decisões por parte das instituições científicas, editores científicos, agências de fomento, mantenedores de repositórios e pesquisadores de modo geral. Acompanhar e compreender os novos modelos de publicação tem importância fundamental para subsidiar o desenvolvimento de novas ferramentas para a publicação e o reuso de dados, de modo a reforçar o sistema de difusão do conhecimento científico de forma aberta e transparente.

2 DIFERENTES ABORDAGENS PARA A PUBLICAÇÃO DE DADOS

No contexto do modelo de publicação científica tradicional, coleções/conjuntos de dados brutos, primários e processados, são desconsiderados como "publicáveis", quando independentes, pelo fato de não conterem inferências, discussões e interpretações dos dados propriamente, tal qual se apresentam nos artigos científicos convencionais (OREGON..., 2017).

Em geral as publicações tradicionais não coadunam com a agenda atual da ciência aberta, que busca maximizar a relação de custo-eficácia dos recursos socioeconômicos investidos em pesquisa, e aumentar a utilidade e aplicação de dados para além do foco ou das limitações de tempo dos coletores de dados originais. Isto porque os artigos de periódicos e de eventos científicos, ainda responsáveis, na maioria das áreas do conhecimento, por espelhar suas inovações e avanços científicos, encapsulam, de forma abreviada, as ideias, análises e pontos de vista, lançados aos dados de pesquisa, dificultando a reprodutibilidade.

Os limites de forma e estrutura do texto científico nesses veículos impossibilita conhecer de forma mais aprofundada os achados da pesquisa, dificultando sua replicação. Tais modelos tradicionais caminham na contramão da tendência atual da ciência que considera dados primários como a moeda mais valiosa da ciência (DAVIS; VICKERY, 2007) e que busca elevar os dados à primeira classe de produtos científicos (CALLAGHAN ., 2013). Dados de pesquisa constituem matérias-primas importantes para a ecologia da ciência e são essenciais para novos ciclos de criação de conhecimento científico, pois fornecem insumos para um processo iterativo no ciclo de vida da investigação, permitindo a continuidade da descoberta científica e da inovação tecnológica.

No atual panorama da ciência aberta, as coleções de dados de pesquisa têm sido elevadas a produtos de pesquisa autônomos e de alto valor intrínseco. Autônomos, pois embora tenham relação com a pesquisa e o contexto por meio dos quais foram originados, estes dados podem ser aplicados a diferentes contextos de modo a responder questões além das propostas e antecipadas pelos investigadores que os coletaram/produziram. O alto valor intrínseco também se relaciona com o potencial de reuso dos dados e com a riqueza de relações que podem ser extraídas de uma mesma coleção de dados (OREGON…,2017).

De modo a potencializar o valor intrínseco e a autonomia dos dados de pesquisa, novos formatos de publicação estão em ascensão e gradativamente conquistaram seus espaços entre os membros da comunidade científica. Com base em Kratz e Strasser (2014), descreveremos três abordagens para a publicação de dados nas seções subsequentes.

2.1 Repositórios de dados

Repositórios de dados são serviços online que podem ser institucionais, temáticos, ligados a comunidades disciplinares, ou a projetos de pesquisa; responsáveis por reunir, descrever, e promover o acesso e a preservação a longo prazo a conjuntos de dados[2].

Os repositórios de dados são parte essencial do ecossistema da publicação de dados, e constituem por si só como uma abordagem de , uma vez que tornam públicas coleções de dados acompanhadas por recursos que otimizem seu potencial de reuso. Embora muitos repositórios de dados tenham surgido nos anos 1960, é notória a recente e vertiginosa expansão destas plataformas, bem como a sofisticação de ferramentas que auxiliem e promovam o reuso efetivo de dados. Pesquisadores contam atualmente com diversas plataformas de repositórios de dados para acesso aberto a dados de pesquisa, possibilitando além de sua preservação a longo prazo, acesso e potencial reuso, funcionalidades adicionais para a manipulação e visualização dos dados, bem como relatórios de estatísticas e métricas de uso. Estes repositórios têm sido populados com dados atendendo às demandas de agências de fomento, instituições de pesquisa e editores científicos que vêm instituindo políticas e mandados de compartilhamento de dados ( ) para pesquisas financiadas com dinheiro público (TENOPIR , 2011).

[2] ou é a agregação de forma visível de dados brutos ou derivados que apresentam uma unidade, alocados conjuntamente, de modo a formar um conjunto coerente, sendo estes potencialmente "citáveis" quando devidamente identificados e representados por metadados (GAILLARD, 2014).

O Re3data.org[3], um projeto que registra repositórios de dados, tem cadastrado atualmente cerca de 1900 repositórios, vinculados a instituições legais de pesquisa, provenientes de diferentes países e nas mais variadas áreas do conhecimento, dedicados ao compartilhamento de dados. Destes, três são mantidos pelo governo brasileiro: o Banco de Dados de Exploração e Produção (BDEP), na grande área de Ciências Naturais, o Repositório de Dados de Levantamentos Biológicos (PPbio) e o recém lançado gerenciado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT).

Diferentes estudos como Candela . (2015) e Wilkinson (2016) apontam barreiras para o reuso dos dados como: o insuficiente contexto e explicações sobre a pesquisa que originou os dados, a ausência de metadados e/ou de validação dos elementos básicos descritivos e identificadores dos dados, sinalizando uma necessidade proeminente de cunho infraestrutural, que estimule e efetivamente permita o reuso de dados científicos.

Parte da solução para contornar esse problema vem de validadores de metadados mais robustos e rigorosos para a descrição das coleções de dados de modo mais completo e detalhado. O repositório de dados em biodiversidade (GBIF)[4], é um exemplo de iniciativa que fornece ferramentas de busca, de visualização e de exportação dos dados, seguindo padrões rigorosos de interno para validação dos dados, e atribuição de licença de reuso para cada conjunto de dados. O GBIF opera sob princípios de endossamento buscando a qualidade dos dados, que eles sejam relevantes ao escopo e objetivos para a comunidade, que a custódia e curadoria dos dados seja estável e persistente, que haja articulação e atuação entre redes nacionais, regionais e temáticas para a publicação e reuso de dados, que os dados possam ser abertamente compartilhados e reutilizados, e que os dados possam ter sua qualidade melhorada pelo autor/publicador com base no de avaliação recebido (GLOBAL…, 2017).

Alguns repositórios de dados atribuem identificadores persistentes aos utilizando o (API) desenvolvido e distribuído pela organização DataCite para atribuição automática do código único identificador O (ORCID) também tem sido utilizado para atribuição e controle de autorias de coleção de dados. Além destes identificadores de objeto e de autor, outros APIs para otimizar a descoberta e a localização de dados, padronizar citações, gerar

[3] www.re3data.org
[4] www.gbif.org

estatísticas de acesso são disponibilizados pela DataCite[5].

O é um exemplo de repositório que integra estes identificadores persistentes, sendo reconhecido na comunidade científica por reunir e gerenciar coleções de dados resultantes de pesquisas a região Ártica, e por oferecer ferramentas avançadas de pesquisa para explorar os dados em profundidade, além de permitir buscas de dados por geolocalização. Outra organização que fomenta a publicação de dados em repositórios é a , uma iniciativa europeia que desempenha várias ações para o gerenciamento de fontes de dados e sustentabilidade dos repositórios de dados, sendo responsável por integrar automaticamente os dados dos projetos europeus e por oferecer mecanismo de busca federada, de modo a facilitar a descoberta de dados.

Ainda que a encontrabilidade e a citabilidade de dados depositados em repositórios venham sendo otimizadas, há de se considerar que esta abordagem de publicação de dados não é muito atrativa do ponto de vista daqueles que compartilham os dados, em termos de recompensa e crédito científico. Sendo assim, outras abordagens para a publicação de dados têm ganhado força mais recentemente, como as publicações ampliadas.

2.2 Publicações ampliadas

Atentos ao valor dos dados da pesquisa, mais intensivamente em 2009, alguns periódicos científicos passaram a solicitar os dados primários como forma de suplementar os manuscritos digitais (CANDELA ., 2015; MOLLOY, 2011). Desde então, uma abordagem para a promoção da ciência aberta e de publicação de dados científicos têm sido as chamadas publicações ampliadas ( ).

A forma como os artigos científicos tradicionais são estruturados e compartimentados apresentam diversas limitações para a transparência e replicação de pesquisas, principalmente nas áreas do conhecimento com foco experimental. De modo geral, os artigos científicos tradicionais tornam visíveis aos pares apenas parte dos dados obtidos na pesquisa e, por vezes, o contexto da pesquisa e seus percursos metodológicos são demasiadamente condensados, fornecendo poucos elementos para se julgar a credibilidade das inferências relatadas nos artigos. Isso se agrava se considerarmos que muitos periódicos adotam o (PDF) como o meio de divulgação dos artigos, o que os tornam ainda mais herméticos, sem interligação

[5] www.datacite.org
[6] arcticdata.io

com recursos complementares que possam auxiliar na interpretação e verificação dos dados relatados na pesquisa.

As publicações ampliadas surgiram pela necessidade de se superar tais limitações, explicitando de forma mais completa e clara a idealização da pesquisa, seus métodos e materiais, bem como o conjunto de dados obtidos no processo de investigação científica. Sobre esse aspecto Sales, Sayão e Souza (2013, np.) esclarecem que:

> […] uma publicação pode ser ampliada a partir da agregação de um ou mais recursos a um e-print. Estes recursos podem ser dados de toda a natureza, outros eprints e metadados e podem ser ainda recursos produzidos ou consultados durante a criação do texto e que, geralmente, apoiam, justificam, ilustram ou esclarecem as afirmações científicas que são apresentadas em uma publicação. [...] Assim, um objeto pode ser parte de um artigo, um , uma imagem, um filme, um comentário, um módulo ou um link para informação em uma base de dados.

A partir da complementação do artigo manuscrito por módulos conectados de arquivos de dados executáveis e interligados, as publicações ampliadas permitem que não só os artigos científicos sejam avaliados de modo mais interativo, mas também potencializam o reuso de dados científicos de modo mais eficiente e eficaz. O artigo científico torna-se um objeto digital mais robusto, tendo caráter agregador de ativos que facilitem a sua interpretação e provejam melhor contextualização acerca do processo de pesquisa, oferecendo contextualização dos dados na própria publicação, mantendo seu sentido original, e possibilitando melhor reutilização para novas pesquisas e reinterpretação dos dados em outros contextos. Além disso, proporcionam maior transparência e possibilidade de verificação dos dados e resultados/análise no momento da leitura, maior interatividade dos métodos de revisão por pares e reduzem o tempo de busca por informações relacionadas à pesquisa em fontes dispersas.

Bardi e Manghi (2014) esclarecem que as publicações ampliadas permitem a difusão e acesso à materiais de pesquisa a partir de ferramentas que incluem desde funções de leitura, descoberta e recomendação baseadas na Web 2.0, ligação com diferentes ativos de pesquisa resultantes ou relacionados ao processo de investigação, e até mesmo ferramentas mais sofisticadas para a execução e plotagem de arquivos digitais complementares com fins de validação, visualização e re-análise dos dados da pesquisa. Em estudo mais recente, Bardi e Manghi (2015) estabelecem uma relação de níveis de sofisticação das funções dos recursos e ferramentas aplicáveis às publicações ampliadas: (1º.) Interligação de ativos externos e de estudos relacionados; (2º.) Capacidades de interação por meio de recursos Web 2.0 pós-publicação (comentários, recomendação, ranking, avaliação); (3º.) Interligação de dados da

própria pesquisa com objetivo de ilustrar e melhor evidenciar as informações; e, (4º.) Possibilidade de reprodução com dados executáveis e recursos de interação para plotagem e visualização de dados. Destes, os dois níveis mais altos são relacionados mais especificamente à publicação de dados, e permitem tanto aos avaliadores, quanto aos leitores, terem mais elementos para embasamento das análises e conclusões extraídas dos dados relatados em um artigo científico.

Um exemplo de periódico ampliado que utiliza dados interligados e requer dados suplementares dos autores é o . De acesso aberto e multidisciplinar, esse periódico oferece recursos que contemplam os quatro níveis citados acima, com vistas a ampliar a interatividade de leitura, a verificabilidade e a transparência em pesquisa. O PLOSOne encoraja o uso de plataformas como o Protocols.io para o depósito e vínculo de protocolos de laboratório, bem como autoriza o uso de uma série de repositórios disciplinares e multidisciplinares (, , , e) para depósito de dados que são articulados aos artigos publicados.

Editores científicos comerciais como a também têm investido mais incisivamente em publicações ampliadas. Alguns exemplos de periódicos que se enquadram na modalidade de publicação ampliada por disponibilizarem , vídeos, modelos gráficos, tabelas e gráficos para download e/ou com recursos de interação vinculados aos artigos são os periódicos , , e .

Cumpre destacar, no entanto, que a gestão de dados quando integrada às publicações ampliadas, tem como desafios oferecer: garantia de persistência e resolução das interligações entre os diferentes recursos de dados (vídeos, imagens, planilhas, modelos 3D, etc.) e o manuscrito digital, interoperabilidade e uso exclusivo de formatos abertos, bem como recursos para citação e mecanismos garantam correta atribuição aos dados interligados. Também há de se considerar que, por se tratarem de materiais suplementares, os dados nessa modalidade de publicação não são necessariamente apreciados pelos pares de forma independente, mas sim dentro do contexto do manuscrito. Desse modo, dentre as diferentes abordagens de publicação de dados que trataremos neste artigo, esta é a abordagem em que os dados de pesquisa ocupam menor posição de protagonismo.

Candela (2015) destacam que as publicações ampliadas podem ser consideradas uma das primeiras tentativas para materializar a publicação de dados científicos mais próxima do modelo de publicação científica tradicional, mas que esse modelo ainda apresenta

desvantagens patentes, pois demanda sofisticada curadoria de ativos científicos dispersos. Acrescido a isto, em geral, a preservação de tais ativos, não permite que os leitores encontrem e liguem dados independentemente do artigo científico. Isso se agrava se considerarmos que muitas publicações ampliadas são pertencentes a editores científicos comerciais, o que torna questionável a garantia de acesso aberto aos dados e de potencial de reuso a longo prazo.

Tendo em vista tais limitações inerentes às publicações ampliadas, um novo modelo de publicação de dados, com maior ênfase no protagonismo destes ativos científicos, e na ampliação das condições de reusabilidade, surgiu como alternativa à comunidade científica: os artigos de dados.

2.3 Artigos de dados

Mais recentemente, a publicação de dados evoluiu para os artigos dedicados exclusivamente para a sua descrição. Essa abordagem surgiu do entendimento de que a publicação de dados seria mais bem aceita e adotada pela comunidade científica, caso espelhasse e preservasse alguns preceitos essenciais do modelo de publicação de científica (CANDELA , 2015), tais como a condição de citação e atribuição de autoria aos criadores e geradores dos dados, e o sistema de avaliação pelos pares.

Os artigos de dados ( ) têm sido considerados parte da solução para melhorar visibilidade dos dados científicos produzidos em pesquisa e aperfeiçoar suas aplicações em contextos diversos. Eles buscam potencializar a descoberta, a visibilidade, a interpretação e a reusabilidade de dados científicos (AUSTIN , 2016). Além disso, de modo geral, os artigos de dados têm sido enaltecidos na literatura como uma alternativa que além de oferecer descrições mais completas dos atributos dos dados, confere legitimidade ao processo de compartilhamento de dados, podendo servir como um mecanismo de recompensa na lógica científica para os produtores destes ativos científicos.

Em termos conceituais os artigos de dados são aqueles buscam descrever uma coleção ou coleções de dados de pesquisa, sem que se estendam à interpretação e inferências dos mesmos. Eles se dedicam à explanação dos métodos para obtenção/coleta dos dados, bem como à descrição da composição da coleção de dados, sua estrutura e formato. Esta modalidade de artigo permite informar a comunidade científica sobre a existência de coleções de dados para potencial reuso e valoriza melhor o conjunto de dados, conferindo-lhe melhor visibilidade, por meio de documentação e metadados mais bem estruturados e rigorosos, que permitem a avaliação das

coleções de dados por pares, e melhor atribuição e reconhecimento aos autores (L'HOSTIS , 2017; AVENTURIER; ALENCAR, 2016).

Chavan e Penev (2011) assinalam três propósitos dos artigos de dados, sendo eles: a) fornecer uma publicação que seja aceita e citável de modo a oferecer crédito acadêmico aos editores de dados; b) descrever os dados de forma estruturada legível por máquinas e também inteligível por humanos e; c) promover e trazer à atenção da comunidade acadêmica a existência de dados de pesquisa possíveis de serem reutilizados.

L´Hostis (2017) esclarecem que os dados nesta modalidade de publicação podem ser integrados aos artigos em tabelas[7], vinculados como material suplementar[8], ou mesmo armazenados em repositórios de dados, tendo apenas o apontamento para a fonte de dados externa (L'HOSTIS , 2017). Quando da terceira opção, em geral, e idealmente, utilizam identificadores únicos, como (DOI)[9] ou um identificador designado por um sistema , além da indicação da referência, com vistas a facilitar a correta citação (KRATZ; STRASSER, 2014). Nesta dinâmica, o artigo de dados cita o conjunto de dados ao qual se relaciona, e o conjunto de dados no repositório também contém a citação do artigo de dados, para que se mantenha a relação de mão dupla.

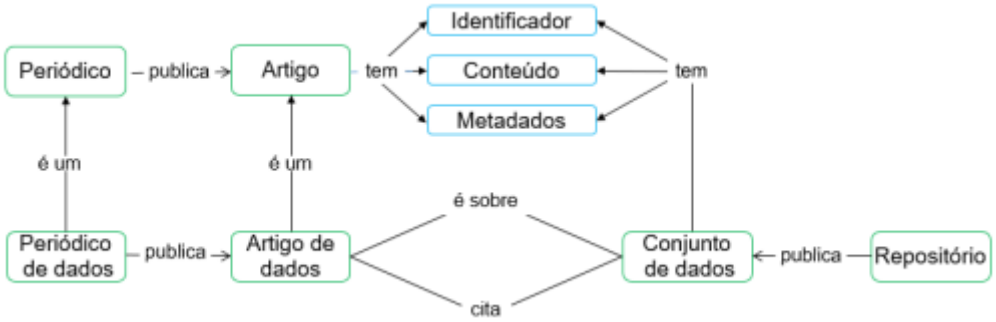
Embora, também possam ser submetidos ao processo de avaliação por pares, ao contrário dos artigos convencionais, os artigos de dados não avançam no processamento dos dados, e não tecem análises e conclusões sobre os mesmos, limitando-se à exposição das circunstâncias de seu processo de coleta (GLOBAL..., 2017). Em outras palavras, enquanto os artigos científicos tradicionais incluem literatura de embasamento e para discussões com vieses analítico e crítico tomando por base os achados da pesquisa, os artigos de dados têm a função exclusiva de relatar as etapas metodológicas para a obtenção dos dados científicos, bem como detalhar os metadados de forma a permitir uma melhor contextualização para reuso futuro e ampliar a reusabilidade dos dados.

Candela (2015) consideram os artigos de dados um tipo de artigo científico, respeitadas as especificidades. Os autores introduzem um mapa conceitual (Figura 1) que traça um paralelo entre estes tipos de artigos, frisando que artigos científicos eletrônicos, possuem identificadores, conteúdo e metadados, enquanto que os artigos de dados possuem

[7] Exemplo: doi.org/10.3897/zookeys.489.9292
[8] Exemplo: doi.org/10.1093/gigascience/gix075
[9] Exemplo: doi.org/10.1002/gdj3.46
[10] Exemplo: doi.org/10.1016/j.dib.2017.07.067

esses mesmos elementos relativos ao conjunto de dados que descrevem. Os conjuntos de dados ( ) devem, por sua vez, estar registrados e hospedados em um repositório de dados e serem referenciados e citados no artigo de dados.

Figura 1: Mapa conceitual acerca dos artigos de dados.



Fonte: Traduzido de Candela *et al.* (2015).

Segundo Callaghan (2013), em geral, a conexão entre o repositório e o artigo de dados que será publicado em um periódico é estabelecida em três etapas. Primeiramente, os autores selecionam um periódico de dados adequado à pesquisa, e verificam quais repositórios são aceitos/autorizados pelos periódicos. Os autores redigem o artigo de dados de acordo com as instruções, modelos e ferramentas recomendadas pelo periódico. Na segunda etapa, os autores submetem o conjunto de dados ao repositório e recebem um identificador e os metadados do artigo, mas não necessariamente disponibilizam os dados abertamente, podendo deixá-los abertos somente ao editor do periódico, para fins de avaliação pelos pares. Os autores então submetem o artigos de dados ao periódico, adicionando o identificador e os metadados providos pelo repositório no ato do arquivamento. Na terceira etapa, o artigo é submetido ao processo de avaliação e revisão pelos pares, sendo que, uma vez aceito, os dados deverão ser disponibilizados de forma aberta, sem restrições de acesso.

Os artigos de dados apresentam várias vantagens para o ecossistema científico, como: a) permitem maior valorização dos dados gerados em pesquisa, por elevarem seu a uma publicação científica legítima e capaz de ser indexada por bases de dados; b) possibilitam a descrição minuciosa dos dados, facilitando a verificação, replicação e reprodutibilidade em pesquisa; c) trazem à tona dados que como materiais suplementares ficam muitas vezes encobertos e são de difícil localização e, d) aumentam o tráfego de acesso a diferentes produções associadas ao conjunto de dados, possibilitando mais citações e descortinando mais possibilidades

para colaboração entre pesquisadores com interesses comuns.

No entanto, os artigos de dados também apresentam alguns inconvenientes. Esta modalidade de publicação de dados demanda tempo e custo dos pesquisadores que tendem a dar prioridade à publicação de diferentes artigos científicos convencionais baseados na mesma coleção de dados, considerando que os mecanismos de recompensa na lógica científica valorizam mais os artigos com viés analítico, do que os artigos descritivos de dados.

Outro inconveniente seria que esta modalidade de publicação não se constitui como uma boa opção para disseminar todos os tipos de dados. Sobre esse aspecto Parsons e Fox (2013) sinalizam que os artigos de dados são adequados para descrever dados relativamente estáveis e em menor escala. Para casos de pesquisas que produzem dados em larga escala, coletados em tempo real e em fluxo contínuo, como em pesquisas climáticas, marinhas e astronômicas, esta modalidade de publicação de dados torna-se menos apropriada.

Um crescente número de editores científicos tem estado atento ao potencial dos artigos de dados para as áreas em que eles são considerados viáveis e, tem buscado minimizar o tempo e esforço envolvido na produção dessa modalidade de publicação, facilitando o processo de geração de por meio de ferramentas automáticas. Alguns periódicos científicos convencionais são receptivos à publicação de artigos de dados, mas periódicos dedicados exclusivamente a essa tipologia de produção científica vêm ganhando espaço no cenário contemporâneo.

Periódicos de dados ( ) geralmente fornecem modelos para descrição e oferecem orientação aos pesquisadores sobre opções de onde depositar e como apresentar e descrever os dados. Estes periódicos contemplam guias próprios para a apresentação e descrição dos dados, podendo adotar critérios genéricos ou específicos de uma área do conhecimento ou disciplina. O repositório de dados nesse caso, pode ser gerenciado pelo próprio periódico de dados, ou mesmo um repositório de dados designado ou autorizado, conforme política editorial do periódico de dados. Alguns mantêm repositórios próprios, enquanto outros apoiam a hiperligação bidirecional entre o artigo de dados e uma coleção de dados hospedada em um repositório de dados externo (AUSTIN , 2016).

No intuito de verificar especificidades de periódicos de dados quanto aos requisitos para

publicação, mapeamos as principais características de oito periódicos exclusivamente dedicados à publicação de artigos de dados indicados em recente relatório sobre dados abertos (BERGHMANS ., 2017), a saber: , , , , , e Scientific Data.

Com base nas informações contidas nos dos periódicos elaboramos o um quadro comparativo[11], que nos permite constatar que todos os oito periódicos são de acesso aberto e seguem o sistema de avaliação por pares como procedimento para a verificação da qualidade, integridade, confiabilidade e consistência dos artigos de dados submetidos. Todos os analisados utilizam a modalidade de licença (CC) BY de atribuição, a licença mais flexível de todas, que permite que outros distribuam, remixem, adaptem e criem outros conteúdos, mesmo para fins comerciais, desde que seja atribuído o devido crédito pela criação original (CREATIVE..., [201?]).

Todos os sete periódicos de dados verificados cobram taxas de processamento ou de publicação com valores distintos. Quanto ao método de arquivamento dos dados, apenas o periódico possui repositório próprio denominado GigaGB para hospedagem dos dados descritos nos artigos, porém o periódico a responsabilidade do autor em obedecer às regras das agências de fomentos e demais instituições às quais estejam subordinados, caso os dados devam ser armazenados outras plataformas. Os demais periódicos referendam opções de repositórios, sendo que alguns flexibilizam essa lista de opções, mediante aprovação prévia do editor. De modo geral, todos os periódicos de dados analisados requerem que as coleções de dados tenham identificadores únicos atribuídos e metadados descritivos informados.

Segundo L'hostis (2017) algumas seções dos artigos de dados são semelhantes a um artigo científico convencional. Isto pôde ser confirmado a partir da observação dos e das instruções para submissão informadas pelos periódicos. É sabido que os artigos de dados não contemplam seções de revisão de literatura, hipóteses e pressupostos, análise e discussão dos dados, e conclusões. Porém, tópicos como resumo, introdução e contextualização, delineamento metodológico e procedimentos, disputa de interesses (em caso de participação da indústria/iniciativa privada), agradecimentos e referências fazem parte da estrutura de artigos de dados.

Em que pesem os tópicos comuns aos artigos científicos tradicionais, há seções peculiares a

[11] Ver: doi.org/10.5281/zenodo.842213

essa modalidade de publicação. Os periódicos de dados requerem uma seção detalhada sobre a descrição dos dados propriamente, incluindo: a composição, o formato, a localização e as formas de acesso e manipulação dos dados, sistemas e software para processamento, entre outros. Há também alguns periódicos que solicitam notas sobre o uso dos dados, isso caso os dados relatados tenham já sido utilizados em outras modalidades de publicação e que indiquem as condições de uso e para usos futuros dos dados documentados no .

Boa parte dos periódicos verificados também solicita que os autores indiquem o potencial de reuso dos dados, articulando como eles podem ser reutilizados dentro e fora do domínio em que foram gerados/coletados, incluindo exemplos de agregação, verificação e replicação e uso dos dados no contexto de ensino de pesquisa.

Berghmans . (2017) e Candela (2015) destacam que os periódicos de dados ainda são um fenômeno em pequena escala e que precisam vencer algumas barreiras e resistências entre os membros da comunidade científica para que atinjam seu verdadeiro potencial. De qualquer modo, os autores reconhecem que este veículo de publicação de dados tem crescido exponencialmente e de forma acelerada nos últimos anos. Para endossar essa afirmação os autores utilizam dados de um estudo bibliométrico que indica o vertiginoso crescimento do acumulado de citações de artigos de dados entre 2012 e 2016, sendo que os periódicos mais prolíficos em número de citações são o , o e, o , dos editores , e , respectivamente.

Embora os estudos citados não esclareçam sobre as possíveis causas para esse aumento de citações, podemos presumir que seja consequência da melhor divulgação desta modalidade de publicação de dados entre os pesquisadores, o que repercute em uma maior visibilidade dos dados existentes e disponíveis com potencial de reuso, bem como pelo fato de que grandes editores científicos, já com prestígio em periódicos científicos convencionais, estão investindo nessa modalidade de publicação, o que vem atraindo e despertando o interesse de cientistas interessados em submeter artigos de dados, e também em utilizar os artigos de dados e os dados como fonte de pesquisa.

De modo a facilitar o processo de produção dos , alguns repositórios de dados estão atuando em colaboração com periódicos de dados para que esse processo seja automatizado. A seguir apresentaremos a abordagem de publicação de dados por intermédio dos repositórios.

3 CONSIDERAÇÕES FINAIS

A publicação de dados tem se estabelecido no meio científico seguindo diferentes abordagens. Apresentamos e organizamos diferentes abordagens identificadas na literatura, seguindo uma coerência de certo modo cronológico, considerando a evolução dos recursos e estratégias para a publicação de dados.

Os repositórios de dados constituem uma modalidade de publicação que tem investido em ferramentas para citabilidade e encontrabilidade dos dados, mas que quando tratamos a publicação de dados, em seu nível elementar, de publicização do registro e da documentação dos dados via repositório, ainda oferecem menos atrativos do ponto de vista da lógica do reconhecimento do trabalho científico, considerando que os pesquisadores tendem a preferir os meios e formatos de disseminação científicos legitimados, que gerem maior visibilidade e que seja convertido em mais créditos e citações.

As publicações ampliadas relacionam coleções de dados, em geral processados, a um artigo científico de abordagem analítica como forma de promover melhor interpretação e verificação no ato da leitura. A publicação ampliada possibilita a interligação entre manuscritos e dados científicos, porém nessa abordagem os dados são acessórios e dependentes do artigo científico. Esse modelo de publicação também não permite a citação de dados de modo independente, dada à inexistência de identificadores e metadados apropriados. Sendo assim, essa abordagem tem seus méritos quanto ao estabelecimento de um formato mais robusto de publicação, mas muitas vezes dificulta o reuso dos dados, tal qual defendido pelo movimento de ciência aberta, principalmente considerando a reusabilidade em contextos diferentes dos antecipados, e em situações transdisciplinares.

Em contrapartida, os artigos de dados elevam os dados científicos à condição de protagonistas, pois se dedicam a descrever exaustivamente a coleção de dados, acompanhados de descrições do contexto, do percurso metodológico e dos aspectos procedimentais da pesquisa, e das possíveis aplicações dos dados. Essa abordagem pode se materializar por meio da publicação dos artigos de dados em periódicos científicos híbridos, receptivos aos , ou periódicos dedicados à publicação de dados. Em geral os artigos de dados são submetidos ao escrutínio dos pares, validando os dados e tornando-os mais transparentes e críveis perante a comunidade científica. Os periódicos de dados também trazem à superfície para comunidades de interesse, coleções de dados com maior potencial de reuso, tendo em vista a garantia de maior

detalhamento documentação, e melhor endossamento. Por preservar características dos
e de cristalizada aceitação e de ampla aquiescência entre os membros da comunidade científica, como , periodicidade e corpo editorial, esta abordagem tem maior potencial de receptividade no meio científico.

No entanto, o processo de elaboração de um requer esforço e tempo que muitos pesquisadores podem não estar dispostos a despender, principalmente se considerarmos a relativa novidade deste tipo de publicação e que muitos pesquisadores podem preferir produzir diferentes artigos científicos com base nos dados coletados, ao invés de publicarem . Por esse motivo, mais recentemente, repositórios e
têm estabelecido parcerias para otimizar e facilitar esse processo por meio de ferramentas automatizadas que geram data papers a partir dos metadados dos dados depositados no repositório. Esses artigos podem então ser submetidos e seguirem o de publicação em um periódico, passando pelo processo de avaliação pelos pares.

Não obstante, os recentes avanços neste novo paradigma de publicação de dados estão ainda em construção e é incerto, o que gera apenas suposições sobre as abordagens possivelmente preferidas e que terão maior adesão pela comunidade científica. O que observamos é que os repositórios de dados desempenham papel central na lógica da publicação de dados, e tem cumprido com diferentes funções, que quando articuladas com periódicos de dados podem aperfeiçoar esse processo. Principalmente quando vinculados a comunidades científicas específicas, a exemplo da biodiversidade, têm investido em ferramentas sofisticadas capazes de aperfeiçoar o da publicação de dados.

Cumpre alertar que muitas publicações ampliadas e , embora
lideradas por grandes editores científicos comerciais. Portanto, devemos considerar o risco futuro da captação dos dados de pesquisas para interesses privados, como quando são publicados exclusivamente como materiais suplementares, vinculados aos artigos, ou depositados em repositórios financiados por esses editores.

Para que a ciência prospere e obedeça aos quatro princípios do FAIR, questões acerca da apropriação de dados devem ser debatidas. Em pesquisa futura, pretendemos relacionar as diferentes abordagens para a publicação de dados aos critérios de encontrabilidade, acessibilidade, interoperabilidade e reuso; com o intuito de classificar o cumprimento de diferentes iniciativas às premissas da ciência aberta.

REFERÊNCIAS

AVENTURIER, P.; ALENCAR, M. C. DE. Os desafios de dados de pesquisa abertos. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, Rio de Janeiro, v. 10, n. 3, p.1-19, 2017.

AUSTIN, C. C. . Key components of data publishing: using current best practices to develop a reference model for data publishing. **International Journal on Digital Libraries**, New York, v. 18, n.2, p.77-92, 2016.

BARDI, A.; MANGHI, P. A framework supporting the shift from traditional digital publications to enhanced publications. **D-Lib Magazine** Reston, v. 21, n.1/2, jan./feb. 2015.

BARDI, A.; MANGHI, P. Enhanced publications: data models and information systems. **LIBER Quarterly**, Munich, v. 23, n. 4, 2014.

BERGHMANS, S. . **Open data:** the researcher perspective. Elsevier: [s.l]. Disponível em: <https://www.elsevier.com/ data/assets/pdf_file/0004/281920/Open-data-report.pdf>. Acesso em: 10 jun. 2017.

CALLAGHAN, S. . **Connecting data repositories and publishers for data publication**. 2013. Disponível em: <http://cedadocs.ceda.ac.uk/id/eprint/951>. Acesso em: 10 jun. 2017.

CANDELA, L. . Data journals: a survey. **Journal of the Association for Information Science and Technology**, New York, v. 66, n. 9, p.1747-1762, set. 2015.
CHAVAN, V.; PENEV, L. The data paper: a mechanism to incentivize data publishing in biodiversity science. , London, v. 12, n. 15, 2011.

CREATIVE COMMONS. **Sobre as licenças**. [201?]. Disponível em: <https://br.creativecommons.org/licencas>. Acesso em: 02 jul. 2017.
CURTY, R. G. **Beyond data thrifting** an investigation of factors influencing research data reuse in the social sciences. 2015. Tese (Doutorado em Information Science and Technology) – Syracuse University, Syracuse. 2015. Disponível em: <http://surface.syr.edu/etd/266>. Acesso em: 18 jul. 2016.

DALRYMPLE, D. Scientific knowledge as a global public good: contributions to innovation and the economy. In: ESANU, J. M.; UHLIR, P. F. (Eds.). **The role of scientific data and information in the public domain:** proceedings of a symposium. National Academy Press: Washington, DC, 2003, p. 35-51, 2003.

EVANS, J. A.; REIMER, J. Open access and global participation in science. **Science**, Washington, DC, n. 323, v. 5917, p. 1025-1025, 2009.

FRY, J., SCHROEDER, R.; DEN BESTEN, M. Open science in e-science: contingency or policy? **Journal of Documentation**, London, v. 65, n.1, p.6-32, 2009.

GAILLARD, R. De **l'Open data à l'open research data:** quelle(s) politique(s) pour les données de recherche? [s.l.]: ENSSIB, 2014.

GLOBAL BIODIVERSITY INFORMATION FACILITY. : how and why you should make biodiversity datasets accessible through GBIF. Disponível em: <http://www.gbif.org/publishing-data/endorsement>. Acesso em: 1 ago. 2017.

KLUMP, J. Data publication in the open access initiative. **Data Science Journal**, Paris, v. 5, n.0, p.79-83, 2006.

KRATZ, J.; STRASSER, C. Data publication consensus and controversies. **F1000Research**, London, v. 3, n. 94, out. 2014.

L'HOSTIS, D. . **Publier un data paper pour valoriser ses données**. 2016. Disponível em: <http://prodinra.inra.fr/record/375633>. Acesso em: 15 jun. 2017

MOLLOY, J. C. The open knowledge foundation: open data means better science. **PLoS Biology, San Francisco**, v. 9, n.12, 2011.

OREGON STATE UNIVERSITY LIBRARIES. **Research Data Services:** data papers & journals. 20 abr. 2017. Disponível em: <