Os princípios FAIR no Ciclo de Vida dos Dados para a Ciência da Informação: maximização do acesso, uso e reúso de dados científicos

Anderson Rafael Castro Simões[a], Renata Lemos dos Anjos[b] e Guilherme Ataíde Dias[c]

1. Introdução

O acesso, uso e reúso dos dados científicos por meio do seu compartilhamento, vêm se destacando na comunidade acadêmico-científica. Pode-se mencionar como exemplo deste fato, as políticas editoriais de vários periódicos científicos eletrônicos, que passaram a solicitar que os dados científicos gerados e/ou coletados no processo de investigação científica sejam devidamente referenciados e fornecidos no momento da submissão do artigo.

De modo semelhante, algumas agências de fomento à pesquisa passaram a solicitar de forma mandatória, que as solicitações de financiamento para projetos de pesquisa incluam um documento conhecido como Plano de Gestão de Dados

a Doutorando em Ciência da Informação pela Universidade Federal da Paraíba – UFPB. E-mail: anderson.simoes@estudantes.ufpb.br. ORCID: https://orcid.org/0000-0002-8268-2783.

b Mestre em Ciência da Informação pela Universidade Federal da Paraíba – UFPB. E-mail: renata.anjos@academico.ufpb.br. ORCID: https://orcid.org/0000-0002-4599-3494.

c Doutor em Ciências da Comunicação/Ciência da Informação pela Universidade de São Paulo – USP. Professor na Universidade Federal da Paraíba – UFPB. E-mail: guilhermeataide@ccsa.ufpb.br. ORCID: https://orcid.org/0000-0001-6576-0017.

(PGD). Este documento planeja e detalha previamente todas as etapas do processo de gestão dos dados ao longo do seu ciclo de vida, essas etapas contemplam os cuidados a longo prazo, contribuindo desta forma para facilitar e simplificar os processos de acesso, uso e reúso a posteriori (SAYÃO; SALES, 2012; WILKINSON et al., 2016; FAPESP, 2017; PLOS MEDICINE EDITORS, 2016; ANJOS; AUTRAN, 2018).

Atualmente, existem inúmeros modelos de ciclo de vida dos dados, elaborados para áreas específicas, com propósitos específicos e, portanto, cada um com suas particularidades. Esses modelos permitem a percepção de etapas cíclicas, características, requisitos e fatores que transpassam essas etapas (SANT'ANA, 2016).

Considerando que a Ciência da Informação pode e deve atuar neste cenário de acesso, uso e reúso de dados, Sant'Ana (2016) propõe o Ciclo de Vida dos Dados para a Ciência da Informação – CVD-CI, composto pelas fases de coleta, armazenamento, recuperação e descarte permeadas por fatores que devem ser considerados em todas elas, como a privacidade, integração, qualidade, direito autoral, disseminação e preservação (SANT'ANA, 2016).

Do mesmo modo, de forma a orientar pesquisadores em todo o processo de gestão, agindo como alicerce para o acesso, uso e reúso desses dados, e de modo a maximizar o valor das publicações acadêmicas que compartilham seus dados, os Princípios FAIR foram criados por partes representantes da

academia, indústria e agências de fomento, em face da necessidade de aprimorar a infraestrutura de suporte para o compartilhamento de dados científicos (WILKINSON et al., 2016).

Estes princípios, cujo acrônimo refere-se a *Findable*, *Interoperable*, *Accessible* e *Reusable*, têm como objetivo fazer com que os dados científicos sejam mais facilmente encontráveis, acessíveis, interoperáveis e reutilizáveis. Ressalta-se que os princípios consideram em suas diretrizes não apenas humanos, mas também máquinas visto que estes dependem cada vez mais do suporte computacional, sobretudo, para lidar com os dados. E referem-se à três entidades, a saber: dados, metadados e infraestrutura (WILKINSON et al., 2016; GO FAIR, online).

De acordo com Jacob (2019), as diretrizes dos princípios FAIR, podem melhorar o ciclo de vida dos dados, sobretudo, usufruindo do uso de tecnologias como, por exemplo, por meio do uso de identificadores persistentes. Nessa perspectiva, o *Consortium of European Social Science Data Archives* – CESSDA, criou um guia denominado "*Data Management Expert Guide*" elaborado por especialistas europeus para auxiliar os pesquisadores das ciências sociais a tornar os seus dados mais "justos", ou seja, FAIR, durante todo o seu ciclo de vida, passando pelo seu planejamento, organização, documentação, processamento, armazenamento e proteção, até o compartilhamento e publicação (CESSDA, 2020). Ainda nessa perspectiva, Miranda (2021) afirma que 90% das diretrizes dos princípios FAIR envolvem

metadados, da mesma forma que, praticamente, os metadados estão presentes em todas as instâncias do ciclo de vida dos dados.

A partir destas breves considerações, este estudo buscou identificar como as diretrizes que compõem os Princípios FAIR podem ser aplicadas nas etapas do Ciclo de Vida dos Dados para a Ciência da Informação (CVD-CI) de maneira a maximizar o acesso, uso e reúso dos dados científicos.

2. Procedimentos metodológicos

Configura-se como uma pesquisa exploratória, bibliográfica, de cunho qualitativo, fazendo uso do método comparativo apontando convergências entre os fatores estudados, no caso entre os Princípios FAIR e o CVD-CI, por meio da inferência da análise de conteúdo (MARKONI; LAKATOS, 2003; BARDIN, 2016).

Inicialmente abordaremos as principais diretrizes propostas pelos princípios FAIR e as fases componentes do CVD-CI. Posteriormente, relacionaremos os Princípios FAIR às fases do CVD-CI, por meio do método comparativo e da análise de conteúdo. Ressalta-se que, neste momento, utilizamos as questões fundamentais que devem ser pensadas no decorrer do CVD-CI propostas por Sant'Ana (2016).

Em seguida, observou-se que os fatores a serem considerados em todas as fases do CVD-CI, também possuem relação com os princípios FAIR, as quais serão elucidadas. Por fim, elaboramos uma figura evidenciando e detalhando as

diretrizes dos Princípios FAIR que podem ser aplicadas no CVD-CI (SANT'ANA, 2016).

3. Resultados e discussões

O primeiro passo para o acesso, uso e reúso dos dados é tornar os dados encontráveis (*Findable*), sendo uma condição para a efetivação dos próximos princípios. Para serem encontráveis, os (meta)dados devem ser atribuídos a um identificador persistente, devem ser descritos com metadados ricos, indexados em recursos pesquisáveis e estar disponíveis online (WILKINSON et al., 2016; MIRANDA, 2021; GO FAIR, online).

Encontrados os dados, faz-se necessário saber como podem ser acessados (*Accessible*), inclui-se aqui o acesso por meio do identificador persistente, acessibilidade dos dados mediante termos de acesso, autenticação e autorização, além da acessibilidade dos metadados quando o acesso livre aos dados não é permitido (WILKINSON et al., 2016; MIRANDA, 2021; GO FAIR, online).

O terceiro princípio (*Interoperable*) ressalta a importância da ligação entre (meta)dados de forma a estabelecer relações semânticas entre os diversos *datasets*, também orienta o uso de formatos abertos e preferenciais, e (meta)dados relevantes por meio da utilização de recursos/instrumentos de controle terminológico como vocabulários controlados que permitem a interoperabilidade entre máquinas (WILKINSON et al., 2016; MIRANDA, 2021; GO FAIR, online).

Por fim, através do quarto princípio (*Reusable*) viabiliza-se o reúso dos dados, permitindo que sejam reutilizados por novos usuários, demandas e aplicações, licenças de uso devem ser explicitamente informadas, assim como questões éticas e legais que regem e protegem os dados. Além do registro, por meio de (meta)dados (relevantes ao domínio para o qual se destinam), sobre o contexto de como os dados foram gerados, incluindo-se questões de proveniência e fluxo de trabalho (WILKINSON et al., 2016; MIRANDA, 2021; GO FAIR, online).

O CVD-CI tem início na fase de Coleta, na qual os dados são obtidos a partir de uma necessidade informacional que determina quais dados serão necessários, como esses dados serão localizados, quais serão os mecanismos para sua obtenção, e as metodologias e ferramentas necessárias para obtenção dos mesmos (SANT'ANA, 2016).

Observa-se que a fase de coleta do CVD-CI associa-se com os quatro princípios FAIR. Com o primeiro princípio *Findable*, quando esse aborda aspectos acerca da encontrabilidade dos dados que facilitam o processo de coleta, por meio da atribuição de um identificador persistente, descrição com metadados ricos que possibilitam a localização dos dados, e a indexação dos mesmos em recursos pesquisáveis e online. Associa-se ao princípio *Accessible* quando a fase de coleta depende do saber de como os dados podem ser acessados, de acordo com os termos de acesso, quais são as questões de autenticação e autorização que regem os dados. O princípio *Interoperable*, associa-se a essa fase,

quando facilita esse momento com o uso de formatos abertos, o uso de recursos/instrumentos de controle terminológico que vão permitir a interoperabilidade entre máquinas, caso a coleta seja realizada por agente(s) computacionais, além de possibilitar a descoberta de outras fontes de coleta visto a vinculação a outros (meta)dados. Por fim, o princípio *Reusable*, rege como os dados coletados poderão ser reutilizados em termos legais e éticos, além de elucidar como esses dados coletados foram gerados por meio de proveniência detalhada.

Quanto aos fatores a serem considerados na coleta, observou-se que os fatores da privacidade e do direito autoral devem ser considerados diante da necessidade em identificar, nas fontes utilizadas para a obtenção dos dados, possíveis aspectos que possam quebrar a privacidade (dos dados pessoais e institucionais obtidos) podendo comprometer as próximas fases do ciclo. Os quais relacionam-se ao princípio *Reusable* que orienta a atribuição de questões éticas e legais na regência de dados sensíveis para que os mesmos sejam reutilizados de forma a respeitar o arcabouço jurídico sustentando este acesso.

Outro fator a ser considerado na coleta, é a integração, na qual os atributos responsáveis pela identificação unívoca de cada registro sejam válidos, sendo relacionada ao *Interoperable*, que recomenda a utilização de recursos/instrumentos de controle terminológico, que vão auxiliar nesse processo de identificação unívoca da integração. O fator da disseminação e qualidade devem ser considerados na fase de coleta quando proporciona

uma maior encontrabilidade e acesso aos dados, relacionando-se ao *Findable*, quando orienta o uso de metadados ricos que devem incluir informações contextuais, sobre a qualidade, condição ou características dos dados. A preservação pode ser considerada quando os dados devem ser preservados e utilizados mesmo com possíveis alterações, que devem ser detalhadas (proveniência detalhada), proposto pelo princípio *Reusable*.

Após a fase de coleta dos dados, percebe-se que os mesmos podem ser posteriormente utilizados em novas análises ou ainda na interação com outros dados. Manifestando-se a necessidade do armazenamento, a próxima fase do CVD-CI. Dessa forma, a fase de armazenamento possui um enfoque tecnológico, que define os aspectos que vão garantir a reutilização desses dados (SANT'ANA, 2016). Ressalta-se que para o CESSDA (2020), a prática do armazenamento e publicação dos dados de maneira adequada (seguindo os princípios FAIR), é o ponto central para tornar os seus dados mais justos, tornar os seus dados mais FAIR.

As diretrizes recomendadas do princípio *Findable* podem ser aplicadas no momento da definição do local de armazenamento (fontes pesquisáveis), e na atribuição dos metadados ricos no momento do armazenamento de forma a facilitar o processo de recuperação (próxima fase). O princípio *Accessible* pode ser aplicado no momento da permissão do acesso apenas aos metadados, em casos específicos (ex. embargo aos dados), e na definição de quais desses dados que foram armazenados podem

ser acessados, e em caso positivo, sob quais termos de acesso e autenticação. O princípio *Interoperable* pode reger a fase de armazenamento no momento da garantia da sua estruturação lógica (uso de recursos/instrumentos de controle terminológico), de como esses dados serão interligados a outros e como essas interligações serão mantidas (referências qualificadas). Por fim, o *Reusable* pode ser aplicado, no que se refere as definições legais e éticas acerca do direito de armazenar esses dados, se os mesmos apresentam riscos a privacidade dos indivíduos ou instituições neles mencionados, além de outros aspectos que garantam sua utilização a posteriori como na garantia de qualidade desses dados com detalhes da proveniência, fluxo de trabalho, e controles de versões.

De forma semelhante a fase de coleta, os fatores da privacidade, direito autoral e qualidade na fase de armazenamento, também se relacionam ao princípio *Reusable*. Na fase de coleta esses fatores devem ser observados no momento em que são coletados, já no momento do armazenamento, estes devem ser observados e realizados de maneira prévia ao armazenamento, sendo definidos para que proporcionem a reutilização dos dados que estão sendo armazenados respeitando as questões éticas e legais referentes a privacidade e direito autoral.

Os fatores da qualidade e preservação também podem ser otimizados por meio do uso de proveniência detalhada, fluxo de trabalho e controle de versões, diretrizes do *Reusable*. Quanto ao

fator da integração, este pode ser otimizado considerando as diretrizes do *Accessible* de como esses dados podem ser acessados (termos de acesso, autenticação e autorização), e do *Interoperable* diante do uso de formatos abertos, na necessidade do uso de recursos/instrumentos de controle terminológicos para identificação e interpretação das entidades e atributos desses dados, no uso de estruturas, padrões de metadados, e na identificação de possíveis relacionamentos, ligações entre dados (referências qualificadas).

A disseminação na fase do armazenamento é fundamental para garantia do acesso futuro, dessa forma, pode ser otimizada com as diretrizes do princípio *Accessible* e *Interoperable* quando deve-se prever meios para que os dados sejam acessíveis, como o uso de identificadores persistentes, descrição com metadados acessíveis, e os termos de acesso, questões de autenticação e autorização, além do uso de recursos/instrumentos de controle terminológicos de forma a garantir a interpretação por máquinas.

Seguindo o CVD-CI, e uma vez que os dados coletados foram armazenados, têm-se a fase da recuperação, que se direciona para a disponibilização desses dados para acesso, uso e reúso, sob o enfoque do responsável pela manutenção desses dados para recuperação (SANT'ANA, 2016).

As diretrizes recomendadas pelo *Findable*, podem ser aplicadas no momento da definição de que, se os dados armazenados serão disponibilizados, aonde e de que forma serão disponibilizados (fontes pesquisáveis), como viabilizar a

encontrabilidade desses dados se por meio do uso de identificadores persistentes e metadados ricos. Quanto ao *Accessible* pode ser aplicado na possível definição de um público-alvo e/ou quem poderá acessá-los mediante autorização e concordância aos termos de acesso. O *Interoperable* pode ter suas diretrizes aplicadas no momento da definição de como dar-se-á a integração entre as estruturas dos dados (padrões de metadados), se haverá a ligação para com outros dados de forma a complementar esse conjunto (referências qualificadas). E o *Reusable* no momento da verificação das questões éticas e legais, sobre o direito de disponibilizar esses dados, considerando riscos à privacidade dos indivíduos e/ou entidades neles mencionados. Bem como, na orientação de proveniência detalhada, informando a frequência de atualização dos dados (controle de versões).

De forma semelhante as fases de coleta e armazenamento, na fase de recuperação, os fatores da privacidade e direito autoral, também se relacionam ao princípio *Reusable*, referentes a sensibilidade dos dados e questões éticas e legais acerca do acesso, uso e reúso, facilitando e assegurando a sua utilização e replicação. As diretrizes do *Interoperable* podem ser aplicadas no fator da integração por meio do uso de padrões metadados, recursos/instrumentos de controle terminológicos, e ligação entre os dados, possibilitando uma análise geral desses dados. A qualidade pode ser otimizada com a aplicação das diretrizes do *Reusable* sobretudo no controle da qualidade, de versões e da proveniência detalhada, bem como do *Findable* e *Accessible*,

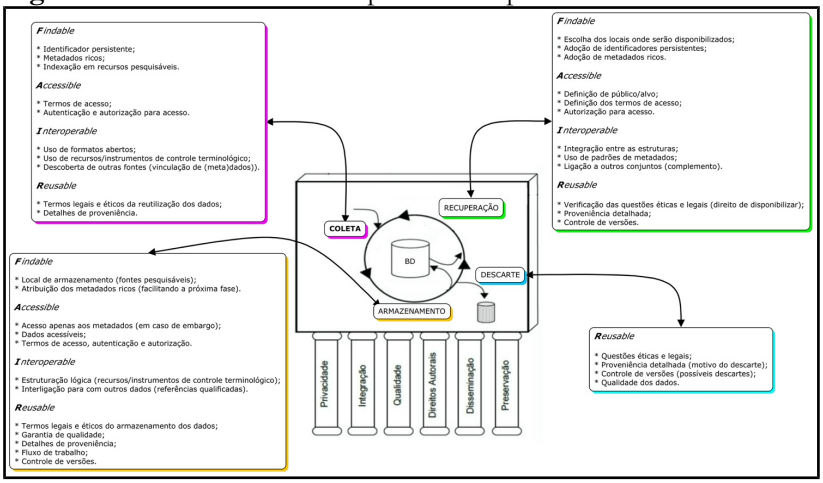
adotando elementos que ampliem a usabilidade e acessibilidade dos dados. De modo semelhante, a disseminação, objetivando a encontrabilidade e (re)uso deve adotar as diretrizes do *Findable* e *Accessible* nas estratégias de localização, de acesso, e sobretudo, no uso de mecanismos para o acesso e a visualização (como formatos e metadados). Por fim, a preservação, atrelada nessa fase ao conceito de interpretação, pode aplicar as diretrizes do *Reusable*, de forma a controlar a qualidade, versões e proveniência, evitando a geração de resultados distintos ao longo do tempo no momento da recuperação.

Finalizando o CVD-CI temos a fase do descarte, consistindo no descarte dos dados julgados desnecessários à determinada pesquisa, ou ainda, aqueles dados que estão acima da capacidade eficiente de tratamento. Nesta última fase, observou-se que apenas o princípio *Reusable* pode ter as suas diretrizes aplicadas, no momento do registro nos controles de versão desse (s) descarte (s), na proveniência detalhada, explicitando o porquê desses dados terem sido descartados, e se, esse descarte afeta de algum modo a interligação a outros dados, ou ainda, a sua encontrabilidade e acesso. Além de dispor do direito, baseando-se em questões éticas e legais, de descarte desses dados (SANT'ANA, 2016).

Quanto aos fatores que devem ser considerados nas fases, a privacidade alega que o indivíduo tem o direito de solicitar que os dados sejam retirados de determinado conjunto, o que pode basear-se nas diretrizes do *Reusable*. O fator da integração,

qualidade, do direito autoral, também podem basear-se no *Reusable* relacionando-se as questões éticas e legais, mas sobretudo, no registro da proveniência de forma a registrar tal descarte, não gerando insegurança legal em trabalhos que já reutilizaram e/ou referenciaram esses dados (dados interligados). O fator da disseminação e preservação podem ser otimizados com o uso do *Accessible* quando os dados descartados ainda podem ser recuperados por meio dos metadados acessíveis, os quais, disporão sobre o descarte dos mesmos.

Figura 1 – Diretrizes dos Princípios FAIR aplicadas no CVD-CI



Fonte: Autor adaptado de Sant'Ana (2016) e GO FAIR (online).

Para melhor visualização das relações e/ou aplicações desenvolveu-se a Figura 1, na qual os princípios foram ligados nas

fases do CVD-CI em que podem ter suas diretrizes aplicadas como forma de otimização do ciclo durante todas as suas fases.

4. Considerações finais

Os princípios FAIR, por meio de diretrizes, objetivam maximizar o compartilhamento e o reúso de dados mais justos, possibilitando, assim, que os mesmos sejam mais encontráveis, acessíveis, interoperáveis e reutilizáveis. De modo semelhante, o ciclo de vida dos dados descreve todos os momentos pelos quais os dados científicos atravessam. No caso do CVD-CI, desde a sua coleta, armazenamento, recuperação, podendo ser uma nova coleta, até o seu descarte.

Como forma de otimizar e maximizar o acesso, uso e reuso dos dados científicos, esta pesquisa em concordância com outras investigações (JACOB, 2019; CESSDA, 2020; MIRANDA, 2021) propôs a adoção das diretrizes recomendadas pelos princípios FAIR no decorrer do ciclo de vida dos dados.

Observou-se que, em sua maioria, os princípios podem ser amplamente adotados no decorrer do Ciclo de Vida dos Dados para a Ciência da Informação (SANT'ANA, 2016), e que essa implementação possibilita a maximização do acesso, uso e reúso dos dados científicos, visto que, enriquece todo o ciclo de vida dos dados, viabilizando todas as suas etapas.

Referências

ANJOS, R. L. dos; AUTRAN, M. M. M. Política de Compartilhamento de Dados Científicos: a adoção nos periódicos

da Ciência da Informação. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 2., 2018, João Pessoa. **Anais** […]. João Pessoa: II WiDat, 2018, p. 269-276. Disponível em: http://www.ufpb.br/widat/contents/documentos/anaiswidat2018.pdf. Acesso em: 19 maio 2021.

CONSORTIUM OF EUROPEAN SOCIAL SCIENCE DATA ARCHIVES – CESSDA. **Data Management Expert Guide**. 2020. Disponível em: https://zenodo.org/record/3820473#.YK_ypZNKhN1. Acesso em: 18 maio 2021.

FUNDAÇÃO DE AMPARO À PESQUISA – FAPESP. Plano de gestão de dados se incorporam a projetos de pesquisa no Brasil. Revista Pesquisa FAPESP, [S. l.], 2017. Disponível: https://revistapesquisa.fapesp.br/planos-de-gestao-de-dados-se-incorporam-a-projetos-de-pesquisa-no-brasil/. Acesso em 21 maio 2021.

GO FAIR. **FAIR Principles**. *Online*. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 20 maio 2021.

JACOB, D. **FAIR principles, an new opportunity to improve the data lifecyle**. 2019. Disponível em: https://hal.archives-ouverEtes.fr/hal-02070883/. Acesso em: 20 maio 2021.

MIRANDA, P. **Como aplicar os princípios FAIR ao ciclo de vida dos dados?** Publicado pelo canal Projeto RCAAP. [S. l.: s. n.], 2021. 1 vídeo (42 minutos e 27 segundos). Disponível em: https://www.youtube.com/watch?v=HZM-JDEaXBc&t=1283s. Acesso em: 21 maio 2021.

PLOS MEDICINE EDITORS. Can data sharing become the path of least resistance?. 2016. **PLoS Medicine.** Disponível em: https://doi.org/10.1371/journal.pmed.1001949. Acesso em: 19 maio 2021.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, [S.l.], v. 21, n. 2, p. 116–142, dez. 2016. ISSN 1981-8920. Disponível em: http://dx.doi.org/10.5433/1981-8920.2016v21n2p116. Acesso em: 19 maio 2021.

SAYÃO, L. F.; SALES, L. F. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação & Sociedade**, v. 22, n. 3, 2012. Disponível em: http://www.periodicos.ufpb.br/index.php/ies/article/view/12224/8586. Acesso em 19 maio 2021.

WILKINSON, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 1, p. 1 – 9, 2016.