SUBSÍDIOS PARA A CONSTRUÇÃO DE UM MODELO DE AVALIAÇÃO DE SISTEMAS DE GESTÃO DE DADOS DE PESQUISA

Luis Fernando Sayão
Comissão Nacional de Energia Nuclear - CNEN/CIN, Brasil
lsayao@cnen.gov.br

Luana Farias Sales
Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT/COEP, Brasil
luanafsales@gmail.com

Resumo: O requisito essencial para uma boa gestão de dados de pesquisa é que ela se realize por meio de plataformas que assegurem infraestrutura tecnológica e gerencial, sustentabilidade econômica e política de longo prazo e que sejam capazes de oferecer serviços disciplinares e um espaço de colaboração para os pesquisadores. Esses sistemas cumprem duplo papel: se configuram como sistemas de informação que apoiam os pesquisadores na publicação, preservação e disseminação de suas coleções de dados; e, ao mesmo tempo, são ferramentas críticas para o descobrimento e o acesso a coleções de dados de outros pesquisadores, possibilitando o reuso e a pesquisa interdisciplinar. Avaliar as plataformas colaborativas de gestão de dados, do ponto de vista do pesquisador que quer depositar seus dados ou encontrar dados de outras fontes, usufruir de serviços informacionais, computacionais; e da instituição de pesquisa que deseja mensurar a conformidade do seu repositório com as melhores práticas, tecnologias, padrões e metodologias, exige considerar parâmetros técnicos, gerenciais e organizacionais. Como contribuição para a composição de um modelo de avaliação de plataformas colaborativas de dados de pesquisa, o presente trabalho apresenta uma sistematização dos principais itens que devem ser considerados na geração de métricas que podem ser aplicadas esses sistemas.

Palavras-chave: Plataforma de Gestão de Dados. Dados de Pesquisa. Avaliação.

SUBSIDIES FOR THE CONSTRUCTION OF AN EVALUATION MODEL OF RESEARCH DATA MANAGEMENT SYSTEMS

Abstract: For the effective management of research datasets, the essential requirement is that it be carried out through platforms that ensure technological and management infrastructure, long-term economic and political sustainability, and that are capable of providing disciplinary services and a space of collaboration for researchers. These systems play two main roles: they are configured as information systems that support researchers in the publication, preservation and dissemination of their own datasets; at the same time, they are critical tools for the discovery and access to datasets of other researchers, enabling reuse and promoting interdisciplinary research. Evaluate data management platforms, both from the point of view of the researcher who wants to deposit their data or find data from other sources, take advantage of informational, computational and training services, as well as from the research institution perspective that wants to measure the conformity of their platform with best practices, technologies, standards and methodologies requires considering a large number of technical, managerial and organizational parameters. As a contribution to the composition of an evaluation model of collaborative research data platforms, the present work presents a systematization of the main concepts that must be considered in the generation of metrics that can be applied to these systems.

Keywords: Data Management Platform. Research Data. Evaluation.

1 INTRODUÇÃO

A ciência contemporânea tem reposicionado os dados de pesquisa para o seu protagonismo histórico, deixando para trás a ideia de mero subproduto das atividades científicas e os qualificando como recurso de primeira grandeza e fonte primária de novos conhecimentos. No cenário atual, parece mais nítida a noção de que o avanço da ciência está proximamente relacionado ao reuso de dados de pesquisa, seja nas ciências exatas, nas ciências sociais e humanas, na arte ou na literatura.

Essa importância dos dados impõe, entretanto, o desafio de gerenciá-los. No passado, o registro dos de experimentos ou de observações astronômicas poderiam estar cuidadosamente preservados em cadernos de laboratório, que atravessavam séculos, e poderiam ser encontrados ainda legíveis no porão de uma antiga biblioteca. Hoje, entretanto, a preservação e a disseminação de dados se desenrolam no mundo digital, efêmero e complexo, cujas ações de gestão devem ser planejadas e intencionais para garantir o acesso e a interpretação dos dados pelo tempo que for necessário.

O crescente interesse pelos dados digitais coletados ou gerados pelas atividades de pesquisa nas últimas décadas criou, por parte dos pesquisadores, instituições acadêmicas, órgão de fomento à pesquisa, uma demanda crescente por estruturas organizacionais, tecnológicas, por capital humano e por políticas públicas que pudessem dar conta da gestão, sustentabilidade e análise desses novos recursos informacionais (SAYÃO, 2017).

Os dados de pesquisa, por conta de sua natureza complexa, heterogênea, moldada pelas idiossincrasias de cada domínio disciplinar e pela dependência de aparatos tecnológicos em constante evolução, dependem de uma forte contextualização para serem interpretados e transmitirem informação e conhecimento ao longo do tempo e, por fim, serem reutilizados por outros pesquisadores. Essas condicionantes implicam na necessidade de uma gestão dinâmica que vai muito além do armazenamento seguro e da disponibilização na *web*. “Um sistema confiável de publicação de dados requer uma efetiva gestão de dados e uma robusta infraestrutura digital” ratifica Claire Austin e seus colaboradores (2015, p. 1). Assim sendo, o potencial de usabilidade de dados de pesquisa que estão sendo compartilhados está fortemente relacionado à adoção de melhores práticas na gestão, na estruturação dos dados, na interoperabilidade, no assinalamento de metadados de qualidade, no licenciamento apropriado e na acessibilidade (AUSTIN *et al.*, 2015). Mas isso não é tudo: as plataformas de dados

devem considerar enfaticamente as condicionantes políticas, legais e éticas associadas ao acesso e reuso dos dados de pesquisa.

Portanto, para apoiar a execução dos processos de gestão e da implantação de serviços de dados voltados para a pesquisa científica é necessário um arcabouço de muitas faces que compreenda todo o ciclo de vida dos dados. Esse ciclo se inicia no planejamento dos dados e continua no seu arquivamento por longo prazo, para as coleções de dados de valor contínuo. No centro desse arcabouço estão os "repositórios online de dados de pesquisa que são grandes infraestruturas de bases de dados desenvolvidas para gerenciar, compartilhar, acessar e arquivar coleções de dados dos pesquisadores" (UZWYSHYN, 2016, p. 1). Mas para cumprirem seus objetivos, os repositórios precisam estar inseridos em ambientes multifacetados que definam seu *modus operandi*.

É necessário, portanto, que os repositórios de dados estejam permeados por políticas organizacionais, condições legais e éticas, processos administrativos, sustentabilidade financeira e temporal e disponham de um elenco de serviços voltados para a sua comunidade-alvo, criando um ambiente multifacetado de gestão de dados e de colaboração entre os pesquisadores. Nessa perspectiva, quando falarmos de repositório de dados, estamos nos referindo a plataformas colaborativas de gestão de dados de pesquisa que incluem processos e informacionais, computacionais em rede, conteúdos distribuídos, processos gerencias e serviços, consubstanciando um ambiente interativo que chamamos de e-infraestrutura de pesquisa. Esses sistemas devem permitir o compartilhamento, interpretação, agregação, análise e síntese dos dados para pesquisadores globalmente dispersos durante o tempo que for necessário.

As plataformas de gestão de dados rapidamente se tornam componentes essenciais da infraestrutura de informação científica mundial. Primordialmente, na forma de grandes bancos de dados hospedados em centros e arquivos de dados disciplinares projetados para a gestão e curadoria. Esses dispositivos são largamente aplicados no âmbito dos domínios científicos que produzem grandes quantidades de dados - identificadas nos segmentos das *Big Sciences* -, tais como Astronomia, Física de Altas Energias, Genética e Ciências Ambientais. Porém, mais recentemente a demanda por repositórios de dados está emergindo no contexto da cauda longa da ciência (SAYÃO; SALES, 2019), ou seja, nos domínios disciplinares em que atividades são desenvolvidas num grande número de laboratórios relativamente pequenos e por pesquisadores individuais que coletivamente produzem a maioria dos resultados

científicos (ASSANTE *et al.*, 2016; HEIDORN, 2008). Nesse contexto, a pesquisa produzida é específica, diversa, inovadora e disruptiva, porém, faltam plataformas disciplinares para a gestão e o compartilhamento dos dados gerados (BORGMAN, 2015).

Na medida em que a gestão de dados de pesquisa se torna rapidamente uma parte importante do fluxo de trabalho dos laboratórios, as instituições de pesquisa começam a desenvolver soluções, definir políticas, estabelecer serviços de interesse de suas comunidades; na outra direção, os pesquisadores precisam identificar repositórios adequados para a publicação dos seus dados – uma exigência cada vez mais presente das agências de fomento, dos editores científicos e da comunidade onde está inserido - e para descobrir coleções de dados de qualidade que possam ser reutilizadas para prosseguimento de seus empreendimentos científicos. Para tal, é necessário refletir sobre quais são os parâmetros – dentre as inúmeras variáveis presentes - que podem dimensionar os requisitos de uma plataforma de gestão de dados.

Tomando como ponto de partida essa questão, o objetivo do presente ensaio é explicitar os principais conceitos que formam o quadro de requisitos que podem contribuir para a composição de diferentes modelos de avaliação para plataformas de gestão de dados de pesquisa. Nessa direção, por meio de pesquisa teórica, bibliográfica e de cunho exploratório, o estudo se vale do discurso dos autores que notadamente se debruçaram sobre a análise das características dessas plataformas e propõe de forma sistemática as possibilidades tecnológicas, padrões e práticas que se desdobram em unidades conceituais, que como um lego, podem se ajustar às diferentes demandas de acesso, arquivamento, preservação, curadoria e demais serviços, e aos diferentes tipos de dados.

2 INFRAESTRUTURA COLABORATIVA DE GESTÃO DE DADOS DE PESQUISA

Infraestrutura é uma noção abrangente e de muitas faces, ela pode ter uma conotação técnica, legal, organizacional e, algumas vezes, cultural e política (GRAFF *et al.*, 2011). Para as fronteiras da pesquisa cientifica, particularmente, todas essas faces se combinam para compor uma e-infraestrutura colaborativa de gestão de dados de pesquisa. O objetivo primordial dessas infraestruturas é a organização, acesso e reuso permanente dos dados e a

interoperabilidade dos sistemas que estão a volta, constituindo uma ecologia de dados. A Partnership for Acessing Data in Europe (PARADE) faz analogia bastante expressiva.

> Dados [de pesquisa] podem ser equiparados ao dinheiro que tem valor unicamente se ele é usado e circula. Como as diferentes moedas que podem ser armazenadas em infraestruturas bancarias globalmente inter-relacionadas, nós precisamos de infraestruturas de dados persistentes, altamente disponíveis e compatíveis, onde os dados de várias disciplinas podem ser armazenados e encontrados (PARADE, 2009).

Alcançar esses patamares de desempenho exige ações de amplo espectro. “Para se constituírem como verdadeiramente úteis, os dados científicos devem possuir estrutura e organização” (RODRIGUES *et al.*, 2010, p. 11), que vão demandar níveis de gestão e curadoria em diferentes escalas. Considerando essa complexidade, o autor e seus colaboradores argumentam que existem duas áreas de requisitos na gestão dados científicos. A primeira está relacionada com as infraestruturas tecnológicas – sistemas, normas e protocolos - necessárias para assegurar a coleta, preservação e acesso aos dados, e ainda a disponibilidade de serviços de amplo espectro. Além do mais, para as infraestruturas informacionais subjacentes às plataformas, a aplicação de padrões técnicos e semânticos, como ontologias, é algo crítico. Mesmo a noção de qualidade é afetada pela proveniência, integridade, autenticidade e propósito dos dados que, por sua vez, têm uma dependência essencial aos esquemas de representação – metadados e documentação. A segunda área de requisitos considera os aspectos políticos, legais e éticos decorrentes do acesso e reuso dos dados além do contexto inicial para que foram gerados, posto que a reutilização é fortemente condicionada por arcabouços legais e éticos e por código de conduta específicos da cultura de cada disciplina, como comissão de ética, políticas de consentimento, propriedade intelectual e tradição de compartilhamento e políticas de incentivo e de financiamento.

O relatório da Knowledge Exchange (GRAFF *et al.*, 2011) identifica quatro elementos-chave para a composição de uma infraestrutura colaborativa de gestão de dados que envolvam todos os atores da comunidade científica: 1) incentivos para o pesquisador na qualidade de produtor dos dados, que incluem reconhecimento, reuso e citação, exigências das agências de fomento e dos periódicos e códigos de conduta disciplinares; 2) capacitação voltada para o pesquisador e também para os cientistas e bibliotecários de dados que apoiem a criação, organização, manipulação, analise e a disponibilidade dos dados para

compartilhamento e reuso; 3) Infraestrutura de dados capaz de apoiar a vasta gama de serviços e 4) financiamento contínuo da infraestrutura.

Esses elementos têm como objetivos estratégicos criar uma ecologia de dados que permita que o compartilhamento de dados seja parte da cultura acadêmica; que a gestão dos dados se torne um componente integral da vida acadêmica profissional; e que a infraestrutura de dados permaneça sólida e sustentável tanto operacionalmente quanto financeiramente.

As infraestruturas de gestão de dados de pesquisa devem, idealmente, estar imbricadas nos arcabouços tecnológicos onde se desenrolam as atividades científicas, que é identificado por muitos autores como uma ciberinfraestrutura colaborativa de pesquisa. Compreende-se esse ambiente como "uma nova forma de cultura científica que se sustenta em uma robusta infraestrutura tecnológica de alto nível" (PÉREZ-GONZÁLES, 2010, p. 3). Idealmente, os procedimentos científicos transcorrem em ambientes colaborativos que integram processos informacionais, computacionais e gerenciais permeados por uma política bem definida que fixa as formas de interlocução técnicas, legais e éticas com todos os stakeholders.

Partindo dos pressupostos acima, definimos nas seções seguintes os componentes e serviços determinantes para a composição de uma plataforma colaborativa de gestão de dados de pesquisa.

3 TIPOS DE PLATAFORMA: A DISCIPLINAR E A MULTIDICIPLINAR

Quando se compara publicações acadêmicas e coleções de dados – ambos produtos de pesquisa – verifica-se que diferentemente das publicações acadêmicas, que são padronizadas transversalmente entre as diversas disciplinas, os dados variam consideravelmente em muitas direções. Isto acontece porque áreas distintas de pesquisa têm diferentes exigências em relação à geração, ao uso e, sobretudo, ao conceito de dado de pesquisa. Essa heterogeneidade intrínseca – estrutural, semântica, conceitual e tecnológica – que caracteriza as coleções de dados é ao mesmo tempo sua riqueza e sua fragilidade. Riqueza na medida em que cria uma ecologia de dados favorável à pesquisa interdisciplinar e à inovação; fragilidade porque é um obstáculo contundente à plena gestão e preservação. Essa condição reflete, porém, a complexidade do ambiente científico e deve condicionar os requisitos de desenvolvimento das plataformas de gestão de dados de pesquisa.

A heterogeneidade dos dados demanda, por exemplo, esquemas de **metadados** que podem variar bastante de domínio para domínio, exigindo modelos de dados flexíveis o bastante para representarem e recuperarem adequadamente os registros de cada área (AMORIM *et al.*, 2015). Nem sempre as plataformas menos especializadas conseguem implementar todo o fluxo da gestão e oferecer serviços talhados às suas respectivas disciplinas, especialmente no que diz respeito a técnicas e ferramentas automatizadas que facilitem a análise e novas explorações de dados (RODRIGUES *et al.*, 2010). É preciso considerar também que as práticas de compartilhamento de dados variam enormemente entre as disciplinas científicas: em algumas áreas, o compartilhamento e o reuso de dados são essenciais para seu desenvolvimento, enquanto noutras, o compartilhamento é quase uma cultura de "troca de presentes", conforme destaca Goodman *et al.* (2014).

De uma forma geral, há uma diversidade de tipos de plataformas que espelham afiliações acadêmicas e institucionais e as políticas e práticas próprias desses segmentos, domínios disciplinares e, sobretudo, a natureza diversificada e heterogênea das coleções de dados de pesquisa. Para a finalidade da presente análise, consideram-se dois tipos: multidisciplinares, chamados algumas vezes de genéricos, e os disciplinares, conhecidos também por temáticos.

- As **plataformas multidisciplinares** gerenciam coleções de dados de diversas áreas, estruturas e tipos, que implicam em representação limitada e serviços básicos. Essas plataformas são essencialmente serviços de compartilhamento e não repositórios de preservação. Elas estão abertas para publicar qualquer tipo de dados, e são especialmente desenvolvidas para dar apoio à publicação de *datasets* produzidos no âmbito da ciência chamada de "cauda longa". São poucos os repositórios multidisciplinares, entretanto é necessário enfatizar que os repositórios institucionais de dados científicos se enquadram, na maioria dos casos, como multidisciplinares posto que têm que abrigar dados de várias disciplinas que são gerados/coletados no âmbito de instituições de pesquisa e ensino de pós-graduação, como as universidades, cujas pesquisas tipicamente se enquadram na cauda longa da ciência (SAYÃO; SALES, 2019)
- Por outro lado, as **plataformas disciplinares** se voltam para domínios específicos, ou para tipos particulares de dados. Em geral possuem modelos de dados

> adequados à representação das coleções e sistemas de gestão voltados para as especificidades de suas comunidades, e oferecem um elenco de serviços mais orientados a essa condição, como curadoria digital, processamento, análises, modelagem, *workflow* e visualização. As plataformas especializadas nem sempre estão baseadas no conceito de repositório, podem estar implementadas como centros de dados, bancos de dados ou arquivos de dados (THE ROYAL SOCIETY, 2012). Vários levantamentos indicam que a maioria das plataformas de gestão de dados se identifica com um domínio particular ou com uma área de estudo bem delimitada (AUSTIN *et al.*, 2015; ASSANTE *et al.*, 2016)

Cada uma das categorias de plataformas, sob a ótica do compartilhamento de dados, tem vantagens e desvantagens: os centros de dados podem não aceitar todos os dados submetidos, considerando que eles aplicam, como os arquivos tradicionais, critérios mais rigorosos de avaliação e seleção de dados para a preservação; por outro lado, os repositórios institucionais podem não ser capazes de apoiar a preservação de longo prazo ou a gestão de dados mais complexos. Baseados no relatório da The Royal Society (2012), relacionamos algumas vantagens de se depositar em plataformas especializadas.

- Oferece uma carteira de serviços especializados;
- Assegura o padrão de qualidade dos dados;
- Possui metodologias de preservação de longo prazo;
- Oferece ferramentas para o armazenamento seguro, controle de acesso e backups regulares;
- Disponibiliza facilidades para buscas precisas e personalizadas e acesso em formatos populares;
- Possibilita o monitoramento do reuso dos dados/gestão de acessos;
- Disponibiliza ferramentas de citação padronizada;
- Promove os dados e incentiva a interação entre pesquisadores;
- Atribui licenças apropriadas aos dados.

Portanto, a escolha por parte do pesquisador para publicação e acesso deve se pautar por plataformas mais próximas de sua área de pesquisa, ou seja, repositórios temáticos que

oferecem esquemas de metadados disciplinares que podem descrever e contextualizar com mais precisão os dados de pesquisa. Porém, se não há repositórios disciplinares para a área de pesquisa dos dados, é melhor que eles sejam depositados em plataformas multidisciplinares como Zenodo[1], Figshare[2], Dryad[3] ou Dataverse[4] , ou ainda no repositório institucional de dados da organização do usuário.

4 SERVIÇOS

O elenco de serviços que deve ser oferecido pelas plataformas colaborativas de gestão de dados de pesquisa deve cobrir todo o leque dos ciclos de vida das coleções de dados – que são específicos, variam em muitas direções e podem perdurar por tempo indeterminado. Como desdobramento dessa heterogeneidade, a oferta de serviços depende de muitas variáveis, como tipo de dados, usuários, domínios disciplinares e seus fluxos de trabalho e ainda das tecnologias disponíveis e do conhecimento por parte dos profissionais envolvidos sobre a geração/coleta dos dados e sobre as etapas de processamento pelo qual eles passaram, e ainda sobre as ferramentas e metodologias aplicáveis. Na presente análise dividimos os serviços em serviços informacionais, serviços computacionais e serviços do pesquisador, ou seja, serviços que têm o apoio significativo dos pesquisadores.

4.1 SERVIÇOS INFORMACIONAIS

O ambiente onde a plataforma está inserida é determinante para a definição do modelo de serviços oferecido aos usuários. Por exemplo, se a plataforma está vinculada a uma biblioteca de pesquisa, serviços biblioteconômicos como serviços de referência, são incorporados ao sistema, como veremos a seguir. A biblioteca de pesquisa tem a capacidade única de romper os silos departamentais de coleções de dados, de agregar recursos e criar uma central de serviços de dados.

[1] Disponível em: https://zenodo.org/.
[2] Disponível em: https://figshare.com/.
[3] Disponível em: https://datadryad.org/.
[4] Disponível em: https://dataverse.org/.

4.1.1 Serviço de referência de dados e consultoria

Podem ser compreendidos como uma extensão dos serviços de referência tradicional da biblioteca de pesquisa que incluem assistência aos usuários para a identificação e recuperação de dados nas várias fontes e diretórios; incluem ainda instruções, cursos e materiais didáticos em torno da descoberta de recursos de dados e de plataformas mais adequadas para a publicação de conjuntos de dados; um item relevante para o pesquisador é o apoio à elaboração do **plano de gestão de dados** de pesquisa, documento que vai se tornando mandatório pelas agências de fomento quando da apresentação de projetos de pesquisa (SAYÃO; SALES, 2015).

4.1.2 Aquisição/desenvolvimento de coleção de dados

Este serviço tipicamente inclui funções como seleção e aquisição de conjunto de dados externos e, principalmente, o apoio ao desenvolvimento de coleções construídas a partir de dados coletados, gerados ou compilados por pesquisadores da instituição.

4.1.3 Competência informacional para pesquisadores

Os pesquisadores não são, via de regra, especialistas em gestão de dados, no entanto cumprem um papel relevante nas etapas preliminares do fluxo de trabalho de tratamento dos dados, principalmente na elaboração do plano de gestão de dados e no **assinalamento de metadados disciplinares** para as coleções de dados geradas/coletadas por suas pesquisas; precisam também identificar repositórios para consulta e depósito de dados. Torna-se importante, portanto, que as plataformas ofereçam cursos, consultorias, publiquem guias e cartilhas e promovam eventos sobre a importância da gestão para o compartilhamento de dados.

4.1.4 Identificações persistentes

A capacidade das coleções de dados hospedadas nos repositórios de serem identificadas e nomeadas permanentemente torna-se essencial para o acesso, preservação e

citação; é um fator importante também nos processos de interoperabilidade e de *linking* com outros recursos via, por exemplo, *linked data*. São muitos os esquemas atualmente em uso, porém o DOI[5] é o padrão recomendado para identificação de coleção de dados. Outros identificadores usados com frequência são o Dspace Handles[6] e o URN[7]; entretanto, muitas plataformas utilizam esquemas próprios de identificação (que prejudicam a universalização e a interoperabilidade). Atualmente, o assinalamento de identificador persistente se torna um fator crítico para a interoperabilidade das plataformas e para a contextualização das coleções de dados.

4.1.5 Citação padronizada dos dados

A citação de dados identifica a prática de atribuir uma referência padronizada a uma determinada coleção de dados. Esta prática tem como propósito descrever os dados permitindo atribuição de crédito aos seus autores, a descoberta e o acesso, e ainda o *link* com outros recursos. Nessa direção, é um mecanismo-chave na publicação de dados de pesquisa, posto que permite que os autores dos dados sejam reconhecidos e que os consumidores possam explicitamente fazer referência aos *datasets* que eles estão reusando em suas pesquisas. De forma geral, a citação de uma coleção de dados apresenta os seguintes elementos: autor, ano de publicação, titulo, editor, nome do repositório e identificador persistente. Há diversas configurações, como, por exemplo, a estabelecida pelo Data Cite[8] que constitui um padrão amplamente adotado. Os serviços de dados têm um papel determinante na citação dos dados, pois podem disponibilizar alguns serviços simples, mas que apoiem significativamente a identificação das coleções de dados.

4.1.6 Controle de versões

Os dados podem não ser produtos de pesquisa finalizados. Nessa perspectiva, a possibilidade das coleções serem editadas/atualizadas por seus autores após serem depositados é um item que deve ser considerado. Embora importante, o controle de versões é

[5] Disponível em: https://www.doi.org/.
[6] Disponível em: https://www.handle.net/.
[7] Disponível em: https://tools.ietf.org/html/rfc8141.
[8] Disponível em: https://www.datacite.org/.

ainda um problema não resolvido plenamente pelas plataformas de dados, o que se observa é que poucas delas oferecem um sistema padronizado de controle de versões. O **carimbo de tempo** (*time stamping*)[9] parece ser a prática mais comum empregada na identificação de arquivos alterados, mas isso não constitui propriamente um controle de versão de conjunto de dados. O Universal Numeric Fingerprint (UNF)[10] é o método sistemático e persistente na identificação de versões de coleções de dados (AUSTIN *et al.*, 2015). É preciso enfatizar que o controle de versões é fator importante na integridade dos dados e para a citação precisa.

4.1.7 Publicação dos dados

A análise da interação do pesquisador *versus* repositório, no evento de publicação de dados na plataforma de gestão, pode ser considerada um momento crítico na aceitação dos sistemas de gestão de dados como meio de arquivamento pelo pesquisador. Esta fase inclui diversas etapas – algumas mais subjetivas e outras quantitativamente mensuráveis – que vão da qualidade ergonômica das interfaces, arquitetura da informação, apoio ao uso de metadados, até o volume e tipo de dados aceitos pelos sistemas e tempo decorrido na submissão, passando por questões de interoperabilidade com outras plataformas, como é o caso do depósito múltiplo interfaceado por mecanismos baseados no protocolo SWORD[11] e a integração com plataformas de publicação acadêmica.

- **Formato de arquivos**

  Os repositórios multidisciplinares geralmente aceitam coleções de dados em qualquer formato, porém os depositantes são frequentemente encorajados a submeter os dados em formatos "amigáveis", ou seja, formatos padronizados, abertos, não proprietários e bem documentados. Algumas plataformas oferecem uma tabela com formatos preferidos para cada tipo de mídia (vídeo, texto, tabelas, imagens etc.), que podem ir de formatos genéricos como pdf e csv à formatos usados em domínios disciplinares específicos como o CSIRO[12], usado comumente em Astronomia. Para o depositante é

[9] Disponível em: https://en.wikipedia.org/wiki/Timestamp.
[10] Disponível em: http://guides.dataverse.org/en/latest/developers/unf/index.html.
[11] Disponível em: http://swordapp.org/.
[12] Disponível em: https://www.csiro.au/en/Research/Astronomy.

fundamental saber que formatos de arquivo o repositório aceita para depósito e se ele precisa converter os seus dados para publicá-los.

- **Tamanho do Arquivo**

  Independente do formato, os repositórios apresentam limitação no tamanho dos arquivos de dados permitido para submissão. Geralmente, ultrapassar esses limites implica em custos adicionais. Por exemplo, o Dryad aceita *uploading* de até 10 GB de materiais para uma única publicação, enquanto o Zenodo atualmente aceita arquivos até 4GB.

- ***Link* com outros recursos**

  O processo de depósito deve dar assistência ao pesquisador para que ele possa efetivamente *linkar* os seus dados a outros recursos, internos e externos à plataforma, como, por exemplo, artigos de periódicos e projetos de pesquisa.

- **Disponibilidade dos dados para a revisão por pares**

  A plataforma precisa oferecer aos revisores oportunidade de consultar os dados como parte do processo de avaliação das publicações consubstanciadas por esses dados.

- **Deposito em outras plataformas**

  É a disponibilidade de dispositivos baseados no protocolo SWORD que permitem depósito em outros repositórios.

4.1.8 Informação de representação: metadados e documentação

Partindo do pressuposto de que a finalidade crítica do processo de publicação de dados é tornar os dados disponíveis para a validação das pesquisas que os geraram e o compartilhamento e reuso por outros atores na área disciplinar onde originalmente foram gerados e coletados e também em outros domínios, fomentando, dessa forma, a pesquisa interdisciplinar, tona-se essencial, portanto, garantir que os dados mantenham as propriedades de serem descobertos, compreendidos e verificados em termos de proveniência e autenticidade ao longo do tempo e do espaço.

Para que isso aconteça, os dados produzidos pelas comunidades científicas devem estar acompanhados por dados auxiliares de representação que forneçam informações contextuais que documentem todas as etapas e processamento pelas quais passaram as coleções de dados e suas características semânticas e estruturais. Vários enfoques são

adotados para contextualizar as coleções de dados por meio de uma documentação apropriada (ASSANTE *et al.*, 2016). A documentação pode estar na forma mais estruturada e padronizada expressa por esquemas de metadados genéricos e descritivos como o Dublin Core (DC)[13], ou disciplinares como o Darwin Core[14] e o Data Documentation Initiative (DDI)[15], que preenchem a necessidade de representação e recuperação de domínios específicos; a documentação inclui também anotações sobre a coleção de dados, cadernos de laboratório e de campo, roteiros de entrevistas, arquivos "leia-me", projeto de pesquisa, entre muitos outros (ROCHA; SALES; SAYÃO, 2017); mais recentemente fazem parte da documentação a descrição das coleções de dados publicada em uma nova concepção de periódico conhecido como *data journals*, que oferece, também, *links* na direção das plataformas onde a coleção completa está depositada.

Há um reconhecimento claro entre os pesquisadores que desejam depositar seus dados em plataformas de gestão de dados de que o apoio à inclusão e à criação de metadados e de outro tipo de descrição e documentação dos *datasets* é um ponto crucial para o processo de publicação de dados (AUSTIN *et al*, 2015), e que quanto mais alta a qualidade dos metadados, maior será a sua capacidade de transmitir conhecimento e de serem descobertos.

Objetivamente, as plataformas precisam (i) dar suporte à aplicação ou ao mapeamento de esquemas padronizados de metadados para a descrição dos dados; (ii) permitir o uso de perfis ou esquemas de metadados customizados; (iii) dar suporte à padrões disciplinares específicos. Alguns sistemas oferecem suporte para a criação de metadados (guias, *templates*, etc.), mas a maioria deixa a questão do controle de qualidade de metadados nas mãos dos provedores de dados.

4.2 SERVIÇOS COMPUTACIONAIS

Serviços que requerem aportes significativos da área de computação e de seus profissionais, incluindo cientistas de dados, programadores, analistas de sistemas e operadores de sistemas de *storage* e pesquisadores.

[13] Disponível em: http://dublincore.org/.
[14] Disponível em: https://dwc.tdwg.org/.
[15] Disponível em: https://www.ddialliance.org/.

4.2.1 Reformatação e limpeza dos dados

Raramente os dados brutos (ou primários) são úteis na forma que são coletados ou gerados por instrumentos ou dispositivos de captura, quase sempre precisam de algum tipo de processamento. Isto porque os dados brutos geralmente não estão no formato adequado para um programador rodar um particular tipo de análise, que impõe a necessidade de reformatação; outro problema recorrente é que os dados brutos frequentemente contêm erros semânticos, dados faltantes ou inconsistentes, portanto é necessária uma "limpeza" antes dos processos de análise (GUO, 2013).

4.2.2 Segurança dos dados

Cobre uma variedade de ações em torno da manutenção da integridade das coleções de dados, significando o estabelecimento de procedimentos físicos e lógicos que impeçam perdas, furtos ou que os dados sejam alterados ou eliminados sem autorização legal, ou mesmo com autorização sem que os eventos relacionados não sejam apropriadamente documentados. Incluem *backups*, arquivamento, proteção física e lógica, criptografia (principalmente nas transmissões) e conformidade com as leis que governam a proteção de dados.

4.2.3 Análise dos dados

Os avanços na computação digital e na capacidade de armazenamento aliados aos novos métodos de comunicação científica, incluindo o uso das mídias sociais, estão introduzindo novos enfoques para as descobertas científicas a partir do acúmulo de dados (SCHMITT *et al.*, 2015). As novas possibilidades de análise devem estar traduzidas em serviços especializados oferecidos pelas plataformas colaborativas de dados

A fase de análise pressupõe um ciclo que inclui: escrever, executar e refinar programas de computadores com o objetivo de analisar e obter novos *insights* a partir das coleções de dados. Os tipos de ferramentas de programação utilizados são geralmente linguagens de *scripts* interpretadas. As preferidas pelos cientistas de dados são *Phyton*, *Perl*,

R e *Matlab*. Entretanto, eles se utilizam também de linguagens compiladas – como C, C++ e Fortran quando for apropriado (SCHMITT *et al.*, 2015).

Segue alguns desses serviços:

- **Visualização de dados** – Dados de pesquisa muitas vezes são difíceis de compreender da forma em que se apresentam, tornando difícil sua interpretação, análise e modelagem. As ferramentas de visualização de dados permitem representar os dados de forma que o seu significado seja comunicado mais claramente, e também a sua relação com outros dados por meio, por exemplo, de representações gráficas.
- **Análise exploratória** – aplicação de uma variedade de metodologias estatísticas avançadas e técnicas de visualização para identificar e validar elementos de dados e/ou determinar se uma hipótese pode ser testada usando os dados.
- **Mineração de dados** – permite gerar múltiplas associações dentro dos limites de uma coleção de dados, que pode render novas informações, especialmente quando combinadas com associações identificadas em outras coleções de dados.
- **Modelagem por computador** – envolve a representação conceitual, matemática, computacional ou física de objetos ou fenômenos do mundo real. É usada para testar hipóteses ou observar e manipular um objeto ou fenômeno que de outra forma seria difícil ou antiético observar e manipular.
- **Simulação interativa e realidade virtual** – uso de computação para criar cenários realísticos onde, de forma diferente da modelagem, o comportamento humano e não um software, guia os resultados da simulação ou da experiência virtual.
- **Workflow científico** – refere-se às etapas abstratas necessárias para completar uma tarefa científica específica, nessa perspectiva os sistemas de fluxo de trabalhos científicos são usados para automatizar etapas tediosas, consumidoras de tempo ou altamente complexas que acontecem nos testes e/ou nas análises experimentais.

4.3 SERVIÇOS REALIZADOS PELO PESQUISADOR

Serviços que são realizados primordialmente pelos próprios pesquisadores por exigir conhecimentos específicos e disciplinares.

4.3.1 Revisão por pares

A revisão por pares de dados depositados não é uma prática universal, entretanto a inclusão de processos de revisão no fluxo de publicação é uma indicação dos padrões do repositório e dimensiona a qualidade geral dos dados, além de creditar confiança nos dados assegurando que um pesquisador possa prosseguir seus estudos baseado nos dados coletados/gerados por outros pesquisadores.

4.3.2 Qualidade dos dados

A gestão da qualidade dos dados é um conjunto de ações que deve permear todo o ciclo de desenvolvimento do projeto de pesquisa. Essas ações asseguram a qualidade dos dados antes deles serem coletados, inseridos no sistema ou analisados. Além do mais, monitoram a qualidade dos dados no decorrer do projeto, aumentando o seu nível de confiabilidade e o seu potencial de uso e compartilhamento.

4.3.3 Gestão dos dados no laboratório

Amorim e seus colaboradores (2015) enfatizam que a maioria das soluções de gestão de dados de pesquisa se concentra na ação dos repositórios de dados, ou seja, na fase final do fluxo de trabalho da pesquisa. Nesta situação há uma lacuna, por parte das plataformas de gestão de dados, no suporte aos estágios preliminares das atividades de pesquisa.

> Introduzir a gestão de dados – e particularmente a produção de metadados – nos estágios iniciais do fluxo de trabalho da pesquisa aumenta a chance de uma coleção de dados alcançar o estágio final desse fluxo, quando ele será arquivado em um ambiente de preservação de longo prazo (AMORIM *et al.*, 2015, p. 109).

Os autores argumentam ainda que registros de metadados melhores e mais detalhados podem ser criados nesse primeiro momento, posto que o ambiente de criação de dados ainda está presente, permitindo um diálogo produtivo entre os pares e um compartilhamento mais imediato. “A coleta de dados é comumente um processo colaborativo, faz sentido, portanto, tornar a produção de metadados também um processo colaborativo” (AMORIM *et al.*, 2015, p. 109).

A capacidade de registrar esses estágios preliminares do fluxo da pesquisa tem sido identificada como um requisito importante por muitas instituições de pesquisa que estão interessadas em integrar aos seus sistemas de gestão de dados soluções que cubram todo o fluxo de trabalho da pesquisa. Considerando que os pesquisadores não são geralmente especialistas em gestão de dados, os ambientes de gestão colaborativa devem ser de uso fácil por não especialistas para que se tornem parte do cotidiano da atividade de pesquisa.

5 INTEROPERABILIDADE

A interoperabilidade como critério de avaliação - no contexto da presente análise - está mais diretamente relacionada à capacidade do modelo do repositório de dados de trocar informações com outros sistemas de forma padronizada, tendo como objetivo mais perceptível o aumento no nível de encontrabilidade dos conteúdos na medida em que eles se tornam disponíveis através de múltiplas rotas. Dessa forma, expor os conteúdos dos repositórios a outras plataformas de pesquisa pode acelerar a visibilidade e o reuso dos dados (AMORIM *et al.*, 2015). Além do mais, a interoperabilidade colabora para que os dados possam ser descobertos e reusados por pesquisadores não pertencentes ao grupo de pesquisa que originalmente os gerou, fomentando a interdisciplinaridade. Não menos importante, os requisitos de interoperabilidade desempenham um papel-chave na contextualização e no potencial de reinterpretação dos dados, na medida em que permitem que eles estejam relacionados – até semanticamente - a outros recursos e atores, como artigos, projetos, pessoas e outras coleções de dados, definindo mais claramente os domínios do seu significado.

O diálogo entre os sistemas de gestão de dados de pesquisa por meio dos mecanismos de interoperabilidade se manifesta de diversas formas. As principais possibilidades de conexão são relacionadas a seguir, e podem ser pautadas como métricas importantes na definição de modelos de avaliação de repositórios de dados.

- **Integração dos sistemas de repositório com os sistemas de publicação**

  Há um consenso perceptível entre pesquisadores e profissionais de informação de que um fator importante para a criação de uma ecologia de dados é a conexão entre os

dados e as publicações que são consubstanciadas por esses dados; é preciso considerar também a vinculação com as publicações que descrevem as coleções de dados, conhecidas como *data jounals*.

- **Coleta automática via Protocolo OAI-PMH**

  O protocolo Open Achives Initiative – Protocol for Metadata Harvesting (OAI-PMH)[16] é um padrão universalmente usado no estabelecimento de interoperabilidade entre repositórios ao mesmo tempo em que também facilita a disseminação de dados; o protocolo é um dispositivo valioso para os *haversters* indexarem os conteúdos dos repositórios.

- **Submissão de dados para múltiplos sistemas via o protocolo SWORD**

  O protocolo Simple Web-service Offering Repository Deposit (SWORD) possibilita a interoperabilidade no momento em que permite o depósito de objetos digitais em múltiplas plataformas e em outros sistemas de forma simultânea.

- **Empacotamento via padrão *Bagit***

  Uso da ferramenta Bagit[17] para empacotamento de dados para depósito em outras plataformas.

- **Acesso aos conteúdos por meio de APIs**

  Oferece acesso alternativo aos dados e metadados por meio de alguma forma de Application Programming Interface (API)[18] para acesso online e intercâmbio.

- **Integração da plataforma de dado com os sistemas de arquivamento confiável**

  Um exemplo importante é a interoperabilidade entre os sistemas de repositório Dataverse e o de preservação Arquivemática[19].

- **Exportação de metadados**

  Algumas plataformas dispõem de facilidades para a exportação de registros que são compatíveis com esquemas de metadados – Dublin Core, MARC-XML etc. (AMORIM *et al.*, 2015); plataformas baseadas em Dspace podem exportar pacotes de informação de preservação (DIP) na forma de registros de metadados METS[20],

[16] Disponível em: https://www.openarchives.org/pmh/.
[17] Disponível em: https://en.wikipedia.org/wiki/BagIt.
[18] Disponível em: https://pt.wikipedia.org/wiki/Interface_de_programação_de_aplicações.
[19] Disponível em: https://www.archivematica.org/pt-br/.
[20] Disponível em: https://www.archivematica.org/pt-br/.

permitindo a ingestão desses pacotes em fluxos de trabalho orientados para preservação de longo termo.

6 REQUISITOS CONJUNTURAIS, POLÍTICOS E ADMINISTRATIVOS

Não obstante as tecnologias computacionais e de redes terem se tornado elementos essenciais na implantação de plataformas de gestão de dados, é necessário considerar que esses dispositivos estão longe de depender unicamente de tecnologias para desempenhar o seu papel como sistema acadêmico de informação. Os parâmetros a seguir conceituam o ambiente político e institucional, bem como as relações de identificação, de organicidade e de visibilidade da plataforma em relação à comunidade científica na qual ela está inserida, além das questões de licença de uso e depósito.

6.1 POLÍTICA DO REPOSITÓRIO

A política de um repositório declara os compromissos que a instituição se obriga em relação aos seus principais *stakeholders* – pesquisadores, curadores, consumidores, financiadores, coletores de metadados entre muitos outros - e em relação às diversas etapas do ciclo de vida das coleções de dados que estão sendo gerenciadas. Esse posicionamento idealmente deve estar manifestado em um documento publicado na página web da plataforma explicitando a política da instituição em relação ao serviço que está sendo disponibilizado. Pode incluir: política de conteúdo, de submissão e de depósito, de direitos autorais, de acesso e reuso da informação, de preservação digital e questões éticas entre outras. A política deve permear todos os processos da gestão de dados, além do mais, ela deve ser um rebatimento harmônico das políticas institucionais e nacionais e das diretrizes internacionais.

É importante, portanto, avaliar se o repositório publica na sua página web um documento que formaliza a sua política; e se os direitos e compromissos dos repositórios e dos depositantes e usuários estão claros, especialmente no que diz respeito ao tratamento dos dados publicados.

6.2 INSTITUCIONALIZAÇÃO

A plataforma deve ser um projeto vinculado a uma ou mais organizações vocacionadas e comprometidas com a gestão e disseminação aberta da informação científica, como são as bibliotecas de pesquisa e centros de dados científicos, ou organizações privadas como os editores científicos que operam repositórios de dados que validam suas publicações, como o repositório Dryad. Nessa direção, um indicador eloquente que deve ser considerado é a informação sobre as agências de apoio à pesquisa ou organizações públicas e privadas que financiam ou dão apoio de outra natureza à plataforma (CLARIVATE ANALYTICS, [201-]).

6.3 RECONHECIMENTO PELA COMUNIDADE CIENTÍFICA

A plataforma precisa ser reconhecido pelas comunidades científicas e possuir uma ligação orgânica com as idiossincrasias da disciplina; precisa do aval dos demais atores, como editores científicos, agregadores e fomentadores de pesquisa como uma fonte de informação confiável, e como parte da infraestrutura informacional voltada para a pesquisa; além do mais, precisa manter um grau de articulação com as demais plataformas da área. O reconhecimento por parte da comunidade de pesquisadores e pelos estudos publicados sobre as plataformas pelos pesquisadores das áreas de ciência da informação, biblioteconomia e computação são indicadores importantes.

6.4 ESTABILIDADE E PERSISTÊNCIA

A estabilidade da plataforma e a persistência dos ativos informacionais que nela são depositados é um item crítico no processo de avaliação. Nessa direção, um repositório de dados deve demonstrar a sua capacidade de permanecer ativo pelo tempo que for necessário e, adicionalmente, apresentar um **plano de sucessão** que indique que a instituição assumirá a custódia dos dados em caso de descontinuidade do repositório. O Data Citation Index[21], por exemplo, adota como critério para indexação da plataforma a verificação regular se novas coleções de dados estão sendo depositadas como forma de verificar se o repositório está correntemente ativo (CLARIVATE ANALYTICS, [201-]).

[21] Disponível em: http://wokinfo.com/products_tools/multidisciplinary/dci/.

6.5 VISIBILIDADE

O relatório da Organização para a Cooperação e Desenvolvimento Econômico (OCDE/OECD) publicado em 2007, intitulado “Princípios e diretrizes para acesso a dados de pesquisa financiados com recursos públicos” declara que a falta de visibilidade de informações sobre coleções de dados de pesquisa coloca um grave obstáculo para o acesso e reuso desses recursos. Isto implica em dizer que as informações sobre as coleções de dados, sobre as organizações produtoras, a documentação sobre os dados e as especificações sobre as condições de uso dessas coleções devem estar disponíveis em escala internacional de forma transparente e preferencialmente via internet (OCDE, 2007).

6.6 PRESENÇA NOS DIRETÓRIOS, AGREGADORES E DISPOSITIVOS DE BUSCA

Em termos práticos, é necessário analisar a visibilidade da plataforma através da sua indexação em diretórios nacionais e internacionais como R3data[22] e o OpenDoar[23]; se o repositório está sendo coletado por agregadores nacionais e internacionais, como o Google Scholar[24] e se há iniciativas institucionais para fomentar a visibilidade do repositório entre as comunidades envolvidas – como parte das ações de curadoria – como, por exemplo, eventos, campanhas, cursos, palestras, publicações e presença na mídia especializada e nas redes sociais.

6.7 LICENÇAS

No contexto dos serviços de dados de pesquisa, uma licença é um instrumento legal através do qual o proprietário de uma coleção de dados publicada estabelece os termos de uso dessa coleção. As licenças associadas ao *datasets* impactam em muitos aspectos o reuso dos dados, regulando desde as atribuições até a exploração comercial. Devem, portanto, ter suas características e diferenças tornadas públicas pelos repositórios.

Quando uma publicação [de dados de pesquisa] ocorre por meio de um serviço de repositório, existem dois tipos de licença envolvidos: a que estabelece o acordo entre o

[22] Disponível em: https://www.re3data.org/.
[23] Disponível em: http://v2.sherpa.ac.uk/opendoar/.
[24] Disponível em: https://scholar.google.com.br/.

repositório e o proprietário dos dados; e a que estabelece o acordo entre o repositório e o consumidor de dados. Ambas as licenças são parcialmente capturadas pelos "termos de serviço" ou pelas "políticas dos repositórios", isto é, elas são partes das regras dos repositórios que os usuários – proprietários dos *datasets* e consumidores – devem concordar em aceitar quando usam os serviços de repositórios. Tanto em termos de depósito quanto em termos de reuso dos dados.

- **Licença de depósito** – especifica os direitos que se espera que sejam outorgados pelo repositório aos proprietários de dados, por exemplo, serviços de curadoria, estatísticas de uso, etc. Em termos práticos, isto significa que o proprietário dos dados precisa verificar se a licença que ele quer associar à coleção de dados é compatível com aquelas suportadas pelo repositório. Em termos gerais, quando o proprietário se registra para depositar a sua *dataset*, ele está aceitando implicitamente as condições e as políticas do repositório. Em relação ao tipo de licença, existe uma tendência, especialmente no contexto das pesquisas financiadas com dinheiro público, de se adotar as licenças estabelecidas – como as licenças Creative Commons - em lugar das licenças proprietárias.
- **Termos de uso** – é parte da documentação do repositório associada com a coleção de dados, pois estabelece as condições sobre como a coleção de dados pode ser realmente usada pelo consumidor após o acesso. A boa prática indica que ela deve estar explicita no momento do *downloading* dos dados e deve ser parte do termo de aceitação dos serviços do repositório
- **Licença de acesso** – em termos práticos, a expectativa sobre os repositórios de dados é que eles ofereçam acesso aos seus ativos informacionais para o público em geral sem a necessidade de registro do usuário, de autorização ou *login*.
- **Tempo de embargo –** é o período de tempo ao longo do qual o acesso à coleção de dados tem o acesso restrito segundo alguns critérios. Por exemplo, estará acessível apenas para revisão por pares. Algumas plataformas, ampliando a ideia de tempo de embargo, oferecem a possibilidade de *upload* de dados de forma que eles sejam mantidos "privados", ou seja, de acesso restrito aos pesquisadores autorizados pelos proprietários da coleção de dados.

- **Licença do material associado ao *dataset*** – a compatibilidade entre as licenças do *dataset* e do material associado a ele, incluindo a documentação, tem uma influência considerável na possibilidade de reuso efetivo dos *datasets* publicados. Nessa direção, esses materiais devem ser publicados com licenças que não invalidem a interpretação e uso dos dados.

As licenças são fatores imprescindíveis para o compartilhamento e reuso e, portanto, são fatores importantes na avaliação de uma plataforma de gestão de dados.

6.8 CERTIFICAÇÃO

A certificação das plataformas de dados por órgãos competentes tem grande importância na medida em que promovem a confiança na usabilidade, sustentabilidade e persistência por longo prazo dos dados disponíveis para compartilhamento. O Data Seal Approval (DAS)[25] confere, por meio de processos de autoavaliação, a certificação básica aos repositórios, e constitui um indicador relevante da qualidade dos services.

7 CURADORIA DE DADOS DE PESQUISA

Neste presente estudo, curadoria de dados é o coletivo de atividade de gestão que circunscreve a adição de valor e de enriquecimento dos dados, bem como a promoção do seu uso. Para tal, a curadoria se inicia ainda no planejamento e criação dos dados e continua até o seu arquivamento em ambientes de preservação confiáveis. A curadoria de dados tem como pressuposto básico assegurar que os dados estejam prontos para os propósitos correntes e futuros, mantendo sua disponibilidade para descoberta e reuso e as condições de proveniência e confiabilidade.

A curadoria de dados de pesquisa pode, de acordo com condições disciplinares, variar bastante, mas inclui a maior parte dos itens relacionados no trabalho, como adição às coleções de dados de metadados, versionamento, identificação persistente, arquivamento. Altos níveis

[25] Disponível em: https://www.datasealofapproval.org/en/.

de curadoria envolvem *links* semânticos com outros materiais publicados, anotações estruturadas baseados em ontologias, entre outros.

8 PRESERVAÇÃO DIGITAL DE LONGO PRAZO

Grande parte dos dados gerados/coletados pela pesquisa contemporânea já está em formatos digitais ou é convertida posteriormente para algum formato digital. Uma parcela desses dados digitais são produtos de experiências que só podem ser reproduzidas a um custo muito alto, ou de observações de fenômenos que não se repetem. Isso implica na necessidade das plataformas de gestão de dados estarem instrumentalizadas com ferramentas, metodologias e expertise para preservação de longo prazo dos dados considerados de valor contínuo e com grande potencial de reuso. É desejável também que a plataforma disponha de uma política explicita de preservação digital, que considere parâmetros arquivísticos tais como proveniência e autenticidade dos dados que não podem ser regerados e estejam conectadas à sistemas confiáveis de preservação fundamentados no modelo de referência ISO/OAIS[26].

9 CUSTO

A operação dos repositórios de dados de pesquisa pressupõe um custo considerável - tanto monetário quanto custos de outra natureza - para as instituições que abrigam estas plataformas. De acordo com Assante *et al* (2016), esse custo está entre os principais fatores que impendem que a publicação de dados de pesquisa seja uma norma corrente na ciência. O custo de publicação pode ser resumido (i) no esforço necessário para preparar os dados de forma que eles possam ser interpretados e usados por outros pesquisadores – o que inclui, por exemplo, documentação; e (ii) no custo monetário de se ter os dados arquivados em ambientes seguros, o que inclui repositórios confiáveis que garantam o acesso por longo prazo aos conteúdos da coleções de dados.

[26] Disponível em: http://www.oais.info/.

Em termos de cobrança, a lógica que se identifica nas principais plataformas, particularmente nas multidisciplinares, deixa claro que os repositórios tendem a cobrar dos proprietários de dados no momento do depósito ao invés do consumidor de dados, que é uma racionalidade mais próxima do acesso aberto. Entretanto, muitos repositórios estabelecem um patamar mínimo em termos de volume de dados em que a submissão é gratuita.

A comunidade científica espera que se estabeleçam novos modelos que reduzam os custos de publicação de dados e que encorajem os pesquisadores a publicarem suas coleções de dados. Por exemplo, nem todos os *datasets* necessitam do mesmo nível de curadoria ou de preservação, portanto é necessária uma especificação mais detalhada dos serviços oferecidos e uma distribuição mais equilibrada de custos; é necessário também envolver as agências de fomento e editores científicos na recuperação dos custos.

10 À GUISA DE CONCLUSÃO

Nos domínios da pesquisa científica contemporânea, o compartilhamento de dados se torna cada vez mais parte da cultura acadêmica, condicionado, muitas vezes, pelas exigências das agências de fomento e dos editores científicos. Como desdobramento direto, a gestão de dados vem se tornando parte integrante da vida profissional dos pesquisadores e da rotina dos sistemas e serviços de informação acadêmicos, principalmente das bibliotecas de pesquisa.

Nesse contexto de mudanças, promover e desenvolver sistemas e competências apropriados para a gestão de dados que estejam alinhados às especificidades disciplinares, e, ao mesmo tempo, às melhores práticas, padrões e exigências internacionais, se tornam um desafio importante para a infraestruturas de informação para a pesquisa em âmbito global.

O ciclo de vida dos dados é longo e guarda muitas especificidades. Ele se inicia antes da geração/coleta dos dados, posto que a sua gênese está na conceituação e planejamento, e continua mesmo depois que as coleções de dados são arquivadas para a preservação de longo prazo em sistemas confiáveis. Fica claro que os procedimentos desse ciclo de vida são mais numerosos e mais complexos do que os procedimentos necessários para a gestão de publicações acadêmicas convencionais, como artigos de periódicos e livros.

Idealmente, esses procedimentos transcorrem em ambientes colaborativos que integram processos informacionais, computacionais e gerenciais permeados por uma política

bem definida que fixa as formas de interlocução técnicas, legais e éticas com todos os *stakeholders*. Isto significa, em termos práticos, adicionar graus de dificuldades à gestão tradicional dos sistemas de informação acadêmicos, incluindo mais recursos financeiros e novas expertises. Para as instituições de pesquisa, significa estar diante de muitas alternativas disponíveis, que podem dificultar a escolha das características mais adequadas aos seus propósitos.

Tomando como ponto de partida essas complexidades, a intenção deste estudo foi alinhar alguns parâmetros necessários ao ajuste dos modelos de avaliação dos sistemas de informação às exigências do protagonismo dos dados de pesquisa, conceituando rapidamente os itens mais importantes, reconhecidos pela literatura, como peças fundamentais para compor os sistemas colaborativos de gestão de dados de pesquisa.

REFERÊNCIAS

AMORIM, R. C. *et al*. A comparative study of platforms for research data management: interoperability, metadata capabilities and integration potential. *In*: ROCHA, Á. *et al*. (ed). **New contributions in information systems and technologies**. Cham: Springer, 2015. p. 101-111.

ASSANTE, M. *et al*. Are scientific data repositories coping with research data publishing?. **Data Science Journal**, v. 15, n. 6, p. 1-24, 2016. Disponível em: http://datascience.codata.org/article/10.5334/dsj-2016-006/. Acesso em: 01 maio 2017.

AUSTIN, C. C. *et al*. Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements. **IASIST Quarterly**, [*S. l.*],p. 24-38, 2015. Disponível em: http://www.iassistdata.org/sites/default/files/vol_39_4_austin.pdf. Acesso em: 01 maio 2017.

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

CLARIVATE ANALYTICS. **The repository selection process**: repository evaluation, selection, and coverage policies for the data citation index within Clarivate analytics Web of Science. [201-]. Disponível em: https://clarivate.com/products/web-of-science/repository-selection-process/. Acesso em: 12 dez. 2018.

GOODMAN, A. *et al*. Ten simple rules for the care and feeding of scientific data. **PLoS Computational Biology,** [*S. l.*], v. 10, n. 4, p. e1003542, 2014. Disponível em: https://doi.org/10.1371/journal.pcbi.1003542. Acesso em: 12 dez. 2018.

GRAFF, M. Van der *et al*. A surfboard for riding the wave: towards a four country action programme on research data. 2011. Disponível em: https://pure.uvt.nl/ws/portalfiles/portal/1427340/Surfboard.pdf. Acesso em: 12 dez. 2018

GUO, P. **Data Science Workflow**: overview and challenges. 2013. Disponível em: https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext. Acesso em: 12 dez. 2018.

HEIDORN, P. B. Shedding light on the dark data in the long tail. **Library Trends**, Champaign, v. 57, n. 2, p. 280-299, Fall 2008. Disponível em: https://www.researchgate.net/publication/49175975_Shedding_Light_on_the_Dark_Data_in_the_Long_Tail_of_Science. Acesso em: 23 out. 2018.

ORGANISATION FOF ECONOMIC CO-OPERATION AND DEVELOPMENT. OECD principles and guidelines for access to research data from public funding. Paris: OECD, 2007. Disponível em: https://www.oecd.org/sti/sci-tech/38500813.pdf. Acesso em: 07 ago. 2018.

PARTNERSHIP FOR ACCESSING DATA IN EUROPE. Strategy for a european data infrastructure: white paper. 2009. Disponível em: https://www.csc.fi/documents/10180/187845/Parade+whitepaper/e0e5c339-1ab5-4724-8d07-6fe8341d1aab. Acesso em: 12 dez. 2018.

PÉREZ-GONZÁLEZ, L. Modelo/s de coste para la preservación de los datos científicos em la e-ciencia. *In*: JORNADAS DE GESTIÓN DE LA INFORMACIÓN, 12., 2010, Madrid. **Anales electrónicos** [...] Madrid: SEDIC, 2010. Disponível em: http://eprints.rclis.org/8555/1/Perez.pdf. Acesso em: 1 maio 2018.

ROCHA, L. de L.; SALES, L. F.; SAYÃO, L. F. Uso de cadernos eletrônicos de laboratório para as práticas de ciência aberta e preservação de dados de pesquisa. **PontodeAcesso,** Salvador, v. 11, n. 3, p. 2-16, 2017. Disponível em: https://portalseer.ufba.br/index.php/revistaici/article/view/24945/15542. Acesso em: 12 dez. 2018.

RODIGUES, E. *et al*. **Os repositórios de dados científicos**: estado da arte. Porto: ACAAP, 2010. Disponível em: https://core.ac.uk/display/55611508. Acesso em: 12 dez. 2018.

THE ROYAL SOCIETY. Science as an open enterprise. London: The Royal Society Science Policy Centre, 2012. Disponível em: https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf. Acesso em: 23 out. 2018.

SAYÃO, L. F. O papel dos repositórios digitais na curadoria de dados. *In*: VECHIATO, F. L. *et al.* (org.). **Repositórios digitais**: teoria e prática. Curitiba: Editora da UTFPR, 2017, v. 1, p. 143-166.

SAYÃO, L. F.; SALES, L. F. A ciência invisível: os dados da cauda longa da pesquisa científica. *In*: DIAS, G. A; OLIVEIRA, B. M. J. F. **Dados científicos**: perspectivas e desafio. João Pessoa: EdUFPB, 2019. No Prelo.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores.** Rio de Janeiro: CNEN, 2015. Disponível em: http://www.cnen.gov.br/images/CIN/PDFs/GUIA_DE_DADOS_DE_PESQUISA.pdf. Acesso em: 07 ago. 2018.

SCHMITT, C. *et al*. Scientific discovery in the era of big data: more than the scientific method. **A RENCI White Paper**, Chapel Hill, v. 3, n. 6, p. 1-22, Nov. 2015. Disponível em: https://renci.org/wp-content/uploads/2015/11/SCi-Discovery-BigData-FINAL-11.23.15.pdf. Acesso em: 12 dez. 2018.

UZWYSHYN, R. Research data repositories: the what, when, why, and how. **Nature,** [*S. l.*], v. 3, n. 3, 2016. Disponível em: http://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml. Acesso em: 12 dez. 2018.