Palabra Clave (La Plata)
ISSN: 1853-9912
palabraclave@fahce.unlp.edu.ar
Universidad Nacional de La Plata
Argentina

# Plataformas de gestão de dados de pesquisa: expandindo o conceito de repositórios de dados

**Sayão, Luís Fernando; Sales, Luana Farias**
**Plataformas de gestão de dados de pesquisa: expandindo o conceito de repositórios de dados**
Palabra Clave (La Plata), vol. 12, núm. 1, e171, 2022
Universidad Nacional de La Plata, Argentina
**Disponible en:** https://www.redalyc.org/articulo.oa?id=350572237003
**DOI:** https://doi.org/10.24215/18539912e171

Artículos de temática libre

# Plataformas de gestão de dados de pesquisa: expandindo o conceito de repositórios de dados

Research data management platforms: expanding the concept of data repositories

*Luís Fernando Sayão*
*Comissão Nacional de Energia Nuclear / Universidade Federal do Rio de Janeiro, Programa de Pós-graduação em Ciência da Informação, Brasil., Brasil*
luis.sayao@cnen.gov.br
https://orcid.org/0000-0002-6970-0553

*Luana Farias Sales*
*Instituto Brasileiro de Informação em Ciência e Tecnologia / Universidade Federal do Rio de Janeiro, Programa de Pós-graduação em Ciência da Informação, Brasil*
luanasales@ibict.br
https://orcid.org/0000-0002-3614-2356

DOI: https://doi.org/10.24215/18539912e171
Redalyc: https://www.redalyc.org/articulo.oa?id=350572237003

Recepción: 02 Junio 2022
Aprobación: 02 Agosto 2022
Publicación: 03 Octubre 2022

## Resumo:

Dados de pesquisa bem gerenciados, no ambiente de pesquisa contemporânea, é reconhecidamente um fator essencial para uma pesquisa de alta qualidade, pois a boa gestão os torna mais fáceis de reuso, o que se traduz em maior coeficiente de colaboração entre cientistas, maximização do retorno do investimento das agências financiadoras de pesquisa, maior transparência nos métodos e fluxos de trabalho, possibilitando a reprodutibilidade dos experimentos científicos. No entanto, a gestão de dados é um problema multifacetado que demanda tecnologias, estruturas organizacionais, conhecimento humano e habilidades para juntar, de maneira complementar, um largo espectro de variáveis, caracterizando-as, dessa forma, como uma equação de resolução complexa. Diante desse desafio, a presente pesquisa parte da seguinte questão: são os repositórios suficientes para solucionar todos os desafios apresentados pela gestão de dados de pesquisa? Para respondê-la foi desenvolvida uma pesquisa de cunho teórico e exploratório, pautada na análise da literatura e na observação de repositórios e plataformas de serviços disponíveis na *web*, culminada no presente ensaio, cujo objetivo é apresentar o conceito de plataforma de gestão de dados de pesquisa, como uma alternativa possível para a resolução de diversos desafios encontrados por pesquisadores e acadêmicos, que visam encontrar, acessar, compartilhar e reusar dados como insumos para novas pesquisas. Conclui-se que a gestão de dados deve se orientar pela oferta de um conjunto de serviços que podem ser classificados como científicos, computacionais, informacionais e administrativos. Esses serviços devem dar suporte próximo aos *workflows* disciplinares, processamento e metodologias de análise por meio de infraestruturas computacionais e informacionais específicas e incorporar expertises multidisciplinares que sejam capazes de lidar com os ambientes e processos tecnologicamente sofisticados da pesquisa atual.

Palavras-chave: Plataformas de gestão dados, Dados de pesquisa, Serviços de gestão de dados, Repositórios de dados de pesquisa.

## Abstract:

In the contemporary research environment, well-managed research data is recognized as an essential factor for high-quality research. This is because good management makes the datasets easier to reuse, which is translated into a higher coefficient of collaboration between scientists, maximizing the return on investment of research funding agencies, increasing transparency in methods and workflows, enabling, in this way, a greater coefficient of reproducibility of scientific experiments. However, data management is a multifaceted problem that demands technologies, organizational structures, human knowledge and skills to combine, in a complementary way, a wide spectrum of variables, thus characterizing them as a complex solution equation.

Faced with this challenge, the present research starts from the following question: are the repositories enough to solve all the challenges presented by research data management? To answer this question, a theoretical and exploratory research was developed, based on literature analysis and observation of repositories and data management service platforms available on the web. As a result, the concept of disciplinary platform for research data management is presented as a possible alternative for solving several challenges encountered by researchers and academics, who aim to find, access, share and reuse data as inputs for new research. It is concluded that the offer of new data management services must be supported by the available computational and informational infrastructures, the analysis methodologies and workflows inherent to the disciplinary research processes and incorporate expertise that is capable of dealing with the environments and technologically sophisticated processes of current research. It is concluded that data management should be guided by the provision of a set of data services that can be classified as scientific, computational, informational and administrative. Those services must closely support disciplinary workflows, processing and analysis methodologies through specific computational and informational infrastructures and incorporate multidisciplinary expertise that can deal with the technologically sophisticated environments and processes of current research.

Keywords: Data management platforms, Research data, Data management services, Research data repository.

## 1. Introdução

Dados de pesquisa frequentemente se manifestam na forma de conjuntos complexos de dados, compostos por diferentes tipos de informação, densamente condicionados por contextos construídos pelas especificidades de seus domínios disciplinares, cujos significados dependem da profundidade das formas de representação de sua cadeia de proveniência. A manutenção desses conjuntos de dados requer conhecimento especializado sobre os ambientes científicos onde são coletados ou gerados e conhecimento avançado em tecnologia computacional e informacional para organizar e arquivar os dados de forma que eles possam ser apropriadamente preservados e reusados (Nielsen & Hjorland, 2014). Na perspectiva de Mayernik e colaboradores (2012), a gestão de dados é um problema multifacetado que demanda tecnologias, estruturas organizacionais, conhecimento humano e habilidades para juntar, de maneira complementar, um largo espectro de variáveis, caracterizando-a, dessa forma, como uma equação de resolução complexa.

O principal objetivo da gestão de dados de pesquisa é revelar o potencial de transmissão de conhecimento dos dados gerados numa investigação científica, transformando o conhecimento, que é local e tácito, em global e explícito para (re)uso no seu percurso espacial e temporal. Isto é realizado por meio de sucessivos graus de agregação de valor que se sucedem por todo o ciclo de vida dos dados – do seu planejamento inicial, ao arquivamento no fim do projeto - que é alcançado por intermédio de processos informacionais, computacionais e científicos. Esses processos que chamamos de serviços de gestão de dados são desenvolvidos no âmbito de arcabouços técnicos, gerenciais e sociais, que no decorrer deste trabalho serão coletivamente denominados de plataforma de gestão de dados de pesquisa.

Uma definição para gestão de dados de pesquisa frequentemente citada e que tem a amplitude conceitual necessária é a colocada por Cox & Pinfield (2014) que, em síntese, preconizam que a gestão de dados de pesquisa é uma série de atividades técnicas e gerenciais associadas ao ciclo de vida dos dados.

> gestão de dados de pesquisa consiste em um número de diferentes atividades e processos associados com o ciclo de vida dos dados, envolvendo o projeto de criação de dados, armazenamento, segurança, preservação, recuperação, compartilhamento e reuso, tudo isso levando em consideração as capacidades técnicas, considerações éticas, questões legais e infraestruturas de governança (Cox & Pinfield, 2014, p. 300).

Essas atividades e processos são exigidos para cobrir um amplo espectro de formas de dados que vão de cálculos em larga escala - originados por dispositivos computacionais de alto desempenho, dados observacionais coletados por instrumentos astronômicos, passando por resultados de experimentos científicos realizados em laboratórios -, até o registro sonoro de entrevistas e a coleta manual de espécimes em um ecossistema. A gestão de dados é, portanto, um conjunto complexo de atividades que envolve uma matriz de desafios técnicos, bem como um grande número de questões culturais, gerenciais, legais e políticas (Pinfield, Cox & Smith, 2014). Com uma longa faixa temporal de aplicação, a gestão efetiva dos dados traz

a promessa de benefícios durante e depois do desenvolvimento de um projeto de pesquisa (Jones, Prior & White, 2013).

O retorno de uma boa gestão de dados [...] são publicações digitais de alta qualidade que facilitam e simplificam os processos em andamento de descoberta, avaliação e reuso em pesquisas subsequentes (Wilkinson *et al.*, 2016, p.1). Nessa perspectiva, a gestão de dados tem como desafio final a otimização do reuso desses dados por seus próprios criadores, por seus pares e ainda por pesquisadores de outras áreas, catalisando, dessa forma, a pesquisa transversal e interdisciplinar – que é onde, via de regra, acontece a inovação. Dados de pesquisa bem gerenciados, no ambiente de pesquisa contemporânea, é reconhecidamente um fator essencial para uma pesquisa de alta qualidade; a boa gestão os torna mais fáceis de usar e reusar, o que se traduz em maior coeficiente de colaboração entre cientistas, maximização do retorno do investimento das agências financiadoras de pesquisa e do atingimento dos objetivos de transparência dos métodos e dos fluxos de trabalho, e o alcance de níveis aceitáveis de reprodutibilidade dos experimentos científicos, paradigma tão caro para a ciência (Strasser, 2015).

Neste sentido, a questão que se coloca é: são os repositórios suficientes para solucionar todos os desafios apresentados pela gestão de dados de pesquisa? A partir desta pergunta, o objetivo do ensaio aqui oferecido é apresentar o conceito de plataforma de gestão de dados de pesquisa, como uma alternativa possível para a resolução de diversos desafios enfrentados por pesquisadores e acadêmicos que visam encontrar, acessar, compartilhar e reusar dados como insumos para novas pesquisas.

O presente artigo é fruto de intensas pesquisas teóricas e empíricas, realizadas durante os anos de 2020 e 2021, que vêm revelando a impossibilidade dos repositórios de dados para solucionar os inúmeros desafios da gestão de dados. A metodologia adotada nesta pesquisa é de cunho teórico e exploratório e se pautou na análise da literatura e na observação de repositórios e plataformas de serviços disponíveis na web. Da observação foi gerada a pergunta que guia esta pesquisa e alguns exemplos que se encontram pelo texto. Já a pesquisa teórica se deu no formato exploratório por ser o objetivo dos autores a proposição de um novo conceito. Por este motivo, a investigação não se valeu de revisão de literatura sistemática e sim do conceito de serendipidade, que proporcionou aos autores a descoberta de relevantes referências para construção pesquisa ao acaso. O estudo das referências aqui citadas embasaram, assim, a proposição do presente conceito.

## 2. Da colocação do problema a uma tentativa de solução: o conceito de plataformas de gestão de dados de pesquisa

Historicamente, grandes partes do esforço no planejamento dos dados e de desenvolvimento de sistemas de gestão de dados ocorreram de forma isolada, escondida por trás das portas dos laboratórios, e com um enfoque comunitário e disciplinar. Esta configuração inicial evoluiu para um cenário que apresenta arquitetura de sistemas, que vão de projetos altamente customizados e de pequena escala, até grandes sistemas de perspectivas mais abrangentes, com alto grau de institucionalização e de internacionalização, e de alcance global. A multiplicidade, diversidade e interoperabilidade das plataformas de gestão de dados põem em pauta o conceito técnico-social de ecossistema de dados de pesquisa, que costura as dinâmicas e interlocuções associadas a esses sistemas pelas pessoas e tecnologias.

De forma ideal, essas plataformas poderiam alternativamente ser criadas em nível nacional ou internacional, onde poderia se esperar uma grande economia de escala, uma centralização de expertises e os serviços não necessitariam ser replicados em inúmeros lugares. O *UK Data Archives*, [1] é um exemplo desse modelo nacional para as ciências sociais no Reino Unido. Para certos tipos importantes de dados e de outros produtos digitais de pesquisa, existem plataformas internacionais com propósitos específicos. Essas plataformas de gestão proporcionam uma curadoria profunda e contínua, um alto grau de integração e uma conexão próxima com as demandas das comunidades disciplinares-alvo, tornando-se, dessa forma, sistemas de referencias para seus respectivos campos de estudos. O *GenBank*, [2] na área de genômica, assim como o

*Protein Data Bank*, [3] e o *UniProt*, [4] são exemplos no escopo das biociências; o *Space Physics Data Facility* (SPDF) [5] e o *Set of Identifications, Measurements and Bibliography for Astronomical Data* (SIMBAD) [6] estão no escopo das ciências espaciais (Wilkinson *et al.* , 2016). Estes sistemas referenciais oferecem dispositivos que assistem aos usuários humanos e máquinas, no acesso aos seus conteúdos de forma dinâmica e precisa, além de proporcionarem uma ampla gama de serviços.

Entretanto, nem todas as disciplinas acadêmicas são cobertas pelos vários centros nacionais e internacionais de dados especializados, atualmente em operação; nem é provável que cada tópico potencial de pesquisa disponha algum dia de uma plataforma específica; além do mais, nem todos os tipos de dados podem ser capturados ou submetidos a essas plataformas, posto que elas geralmente interpõem vários níveis de exigências para a publicação de dados. Todavia, muitos *datasets* importantes emergem de pesquisas tradicionais realizadas nas bancadas dos laboratórios e não se ajustam aos modelos de dados das plataformas de propósitos temáticos existentes e às barreiras interpostas. Nada obstante, esses conjuntos de dados não são menos importantes em relação à integralidade e à reprodutibilidade da pesquisa e às possibilidades de reuso (Wilkinson *et al.*, 2016), sendo assim, eles precisam ser gerenciados.

Portanto, neste cenário multifacetado, é preciso considerar que existem muitos pequenos grupos de pesquisa ou mesmo pesquisadores individuais, localizados na distribuição estatística conhecida como "cauda longa da pesquisa" (Sales & Sayão, 2018), que trabalham em diversos campos produzindo dados com características muito específicas e que têm requisitos que não são facilmente generalizáveis; ou áreas disciplinares que são tão estreitas para justificar o custo de se estabelecer e manter grandes centros de dados. Além disso, há as universidades, centros de pesquisa e outras organizações produtoras de conhecimento científico que desejam integrar suas coleções de dados às suas memórias acadêmicas por meio de plataformas de gestão de dados, desenvolvidas em torno de repositórios institucionais.

Aparentemente, em resposta a essa demanda, vão surgindo inúmeros repositórios multidisciplinares e de múltiplos propósitos, numa escala que vai de repositórios institucionais, por exemplo, pertencentes a uma única universidade, à repositórios abertos de escopo global, tais como *FigShare, Dryad, Mendeley Data, Zenodo, DataHub, DANS e EUDat, entre outros*. Estes repositórios aceitam um amplo espectro de tipos de dados que variam em termos de formatos, volume, modelos e estruturas. Observa-se também que eles não tentam integrar ou harmonizar os dados depositados e interpõem poucas restrições aos metadados assinalados na publicação dos dados. O ecossistema de dados resultante, portanto, parece afastar-se da tendência relacional e está se tornando mais diverso e menos integrado, exacerbando, como consequência, os problemas de descoberta e reusabilidade para seres humanos, e muito mais para stakeholds computacionais "Não obstante, são precisamente os tipos de análise integrativa, profunda e ampla que constituem a maior parte da *eScience*", concluem os autores (Wilkinson *et al.*, 2016, p. 3).

Dito de outra maneira, o investimento na construção de ambientes tecnológicos para gestão de dados de pesquisa vem colocando em pauta outro desafio que é a oferta de serviços úteis que possam apoiar a *eScience* por todo o seu processo e não apenas ao final quando a pesquisa é finalizada e os dados devem ser depositados. Tudo isso somado coloca em pauta o papel dos repositórios como solução singular para a gestão de dados e nos permite a proposição de um conceito de plataforma de gestão de dados de pesquisa, como uma proposta de ferramental em que as atividades, os serviços e os processos que compõem a gestão de dados se agregam e se complementam num ambiente, compreendido como um arcabouço técnico, social e gerencial, em que se efetivam os cuidados com os dados, segundo políticas e diretrizes institucionais definidas para tal. A seção a seguir delineará os contornos de serviços disciplinares de gestão de dados que possam ser oferecidos por meio dessas plataformas.

## 3. Delineando os contornos dos serviços de gestão de dados

Como ratificado por Jones, Prior & White (2013, p. 5), "para dar apoio efetivo à gestão e ao compartilhamento de dados, uma instituição necessita de uma estratégia coerente e de um conjunto de serviços". Mas o que poderia significar este conjunto de serviços de gestão de dados? Naturalmente ele tem um espectro contínuo que varia em termos disciplinares, cultural e epistemológico, institucional e político, e ainda depende das bases tecnológicas disponíveis para a gestão de dados. De fato, as instituições de pesquisa podem oferecer serviços de dados numa grande multiplicidade, que varia não somente nos tipos de serviço, mas também na profundidade e alcance em que esses serviços são disponibilizados, nos níveis de especificidade e comprometimento e para quem e com que objetivos esses serviços são oferecidos (Choudhury *et al*., 2018).

Para exemplificar, Fearon Jr., Gunia, Lake, Pralle & Sallans (2013) apontam que serviços de gestão de dados englobam o fornecimento de informações, consultoria, treinamento e ainda o envolvimento ativo no planejamento da gestão de dados, orientação durante a pesquisa (por exemplo, aconselhamento sobre o armazenamento de dados e segurança de arquivos), documentação e metadados, compartilhamento de dados de pesquisa e curadoria (seleção, preservação, arquivamento, citação) de projetos concluídos e dados publicados. Já sob a perspectiva de Choudhury e colaboradores (2018), os serviços de gestão de dados incluem a oferta de infraestrutura necessária para realizar a curadoria de dados por meio de licenças para preservação, análises e ferramentas de acesso; a disponibilidade de espaço em sistemas de armazenamento financiados pela organização para dados curados; treinamento e consultoria que permitam o pesquisador explorar os serviços de dados oferecidos pelas várias unidades da instituição. Complementando, Tang & Hu (2019) apontam que no diagrama de componentes de gestão de dados de pesquisa, as atividades abrangentes incluem "política e estratégias de gestão de dados" e "plano de negócios e sustentabilidade". Subjacente ao estabelecimento de serviço de gestão de dados de pesquisa, vários níveis de orientação, treinamento e suporte são necessários. Para esses autores, o ponto focal do processo de gestão de dados deve dar proeminência aos componentes de serviço de gestão de planejamento, gerenciamento de dados ativos, seleção e compartilhamento, bem como repositórios e catálogos de dados. Neste sentido, os repositórios ou os catálogos de dados são apenas mais um serviço dentre inúmeros outros que uma plataforma de serviço de gestão de dados pode oferecer.

É importante observar que são muitas as diferenças entre a gestão de recursos mais tradicionais e o nível de exigências técnicas e de infraestruturas e expertises necessárias à gestão de dados de pesquisa. Um livro, por exemplo, tem uma catalogação universal e padronizada, as diferenças de tratamento entre disciplinas são poucas e seus processos estão focados na pós-publicação; o mesmo não se pode dizer de dados de pesquisa e de outros objetos digitais de pesquisa, como base de dados e códigos, cuja gestão tem que se preocupar com o longo e idiossincrático ciclo de vida que se inicia ainda na fase de planejamento - muito antes da publicação e arquivamento, indo até a pós-publicação, mas num processo ainda mais complexo do que era executado na gestão das publicações bibliográficas. Some-se a isso, toda a peculiaridade própria que exige a articulação da gestão com o ciclo de vida do projeto de pesquisa. Neste contexto, o que se observa é que "O leque de competências e conhecimentos necessários para entregar serviços de gestão de dados é ditado em grande parte pelas fases individuais do ciclo de vida do projeto", confirmam Jones, Prior & White (2013, p. 3). Assim, a escala de serviços que as instituições de pesquisa oferecem pode variar não apenas nos tipos de serviços disponibilizados, mas também no nível de profundidade em que eles atuam, e no universo de usuários para quem os serviços são oferecidos (Choudhury *et al.*, 2018). Pesquisadores, professores e estudantes de pós-graduação são os clientes-alvo mais prováveis dos sistemas de gestão de dados, porém outros stakeholders devem ser considerados, como os gestores de C & T, financiadores e comunidades especificas – como engenheiros e agrônomos -, que reusam os dados, especialmente os dados com alto grau de processamento, nos seus projetos e empreendimento, como na construção das fundações de uma usina nuclear ou na seleção de cultivares. Os serviços podem estar distribuídos por várias unidades da instituição ou concentrados e coordenados por uma unidade, possivelmente a biblioteca de pesquisa.

A visão fragmentada e heterogênea sobre os serviços de gestão de dados – que por fim reflete as múltiplas faces da atividade de pesquisa - cria um obstáculo no delineamento dos seus contornos e na enumeração do diagrama dos seus componentes. Por este motivo, conhecer as infraestruturas necessárias para estruturação de plataformas de gestão de dados é uma condição urgente para aqueles que acreditam que é necessário mais do que a construção de um espaço de armazenamento para atender as necessidades reais dos pesquisadores durante todo o desenvolvimento da pesquisa.

## 4. Infraestruturas necessárias para a estruturação de plataformas de gestão de dados de pesquisa

Infraestrutura é uma noção de grande amplitude e multidimensional. Ela pode ter uma conotação técnica, legal, organizacional e, em muitos casos, é imprescindível considerar também os aspectos sociais, culturais e políticos. De fato, é assim no domínio da ciência: o projeto de infraestrutura de pesquisa é simultaneamente uma questão tecnológica, uma questão de identificação das necessidades da pesquisa em áreas disciplinares específicas e uma questão política. Essa ótica mais geral se aplica às infraestruturas institucionais de gestão de dados de pesquisa que precisam oferecer tecnologias e ferramental, processos, políticas, recursos e treinamento para os vários e diversificados estágios da gestão de dados.

De fato, da mesma forma que as instituições devem providenciar infraestruturas básicas para a pesquisa – tais como, laboratórios, instrumentação, computação de alto desempenho, redes, reagentes e muito mais – elas devem também tomar medidas para uma gestão adequada dos dados. Isto pressupõe um amplo espectro de atividades gerenciais, tecnológicas e informacionais que inclui profissionais de informação treinados para apoiar pesquisadores no planejamento e gestão de seus dados, no acesso a dispositivos de armazenamento seguro e backups durante o desenvolvimento do projeto e disponibilidade de plataformas de acesso e de preservação de longo prazo, necessárias após o fim da pesquisa (Strasser, 2015). É imprescindível também um corpo de normas, padrões e boas práticas que permitam, principalmente, uma interlocução em níveis variados dos sistemas e serviços, tanto local quanto global, que pode ser traduzida por interoperabilidade. Nesta categoria, à guisa de exemplo, estão os **padrões de modelo de dados** - geralmente estabelecidos por um domínio disciplinar ou repositório – que determinam a estrutura dos vários componentes de uma coleção de dados, que, por fim, têm efeito sobre as interfaces de interação com os usuários humanos e computacionais e sobre os níveis de interoperabilidade do *dataset* (Choudhury *et al.*, 2018).

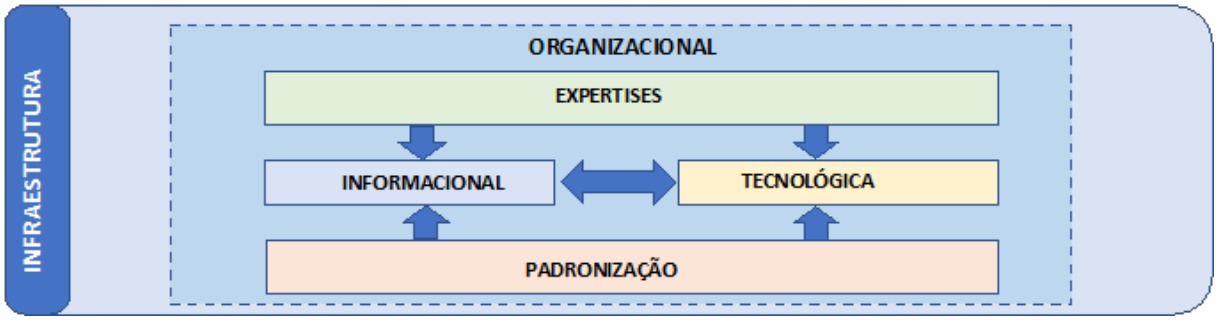
Quando comparamos a publicação acadêmica tradicional com a publicação de dados, verificamos que as infraestruturas subjacentes à publicação acadêmica criam uma ponte epistemológica entre disciplinas, tendo como ponto agregador as bibliotecas de pesquisa, que selecionam, coletam, organizam e tornam acessíveis publicações de todo o tipo e de todas as áreas. Por sua natureza, as instituições sociais trabalham para estabilizar práticas particulares e formas de conhecimentos. Em certo sentido, as instituições são infraestruturas sociais em si mesmas. Nessa direção, as infraestruturas técnicas estão entrelaçadas com as infraestruturas sociais das instituições, muitas vezes mediadas por padrões, protocolos, documentos e artefatos que ligam os aspectos sociais e técnicos das infraestruturas (Leonardi, 2010). Entretanto, não existe ainda infraestrutura dessa magnitude para os dados. Algumas poucas áreas têm mecanismos consolidados para publicar dados; outras estão nos estágios de desenvolvimento de padrões e práticas para agregar seus dados e torná-los amplamente acessíveis. Um problema-chave nas instituições de pesquisa, como observam Mayernik e seus colaboradores (2012, p. 158), "é a falta de uma infraestrutura confiável que possa ser implantada num nível institucional", essa "falta de infraestrutura para dados amplifica a descontinuidade na publicação acadêmica", acrescenta Borgman (2007, p. 155).

Os arcabouços infraestruturais voltados para a gestão de dados são diversos e fragmentados em termos de fluxos, complexidade, aplicação e topologia, e organizados de forma diferente pelas várias disciplinas e em diferentes países (Graaf & Waaijers, 2011). Contudo, crescentemente, as infraestruturas moldam os padrões

e as práticas da gestão de dados. Diante desse fato, o conhecimento sobre a origem, domínio disciplinar, grau de processamento, sistemas de coleta, *workflows* etc. parece ser de importância crítica na concepção de infraestruturas voltadas para a gestão de dados (Sayão & Sales, 2020).

Como pode ser visto na Figura 1 a seguir, consideramos cinco instâncias de infraestruturas necessárias à implantação de sistemas de gestão de dados: Instância de padronização, Instância tecnológica, Instância informacional, Instância profissional e Instância organizacional.

FIGURA 1
Instâncias de infraestrutura necessárias às plataformas de gestão de dados.




Fonte: elaboração própia.

- **Instância de padronização -** Normas e padrões são formas consensuais de codificar o conhecimento que circula transversalmente por comunidades para assegurar uniformidade e similitude aos seus produtos e processos através do tempo e do espaço. Eles refletem o conhecimento mais atual sobre as práticas profissionais e aumentam a interoperabilidade, a consistência, a preservação, a reusabilidade, a segurança e a proteção das coleções digitais. Portanto, assegurar que em um ecossistema científico, em que as infraestruturas estão globalmente dispersas, seus produtos se alinhem aos Princípios FAIR *F(findable), A(acessible), I(Interoperable) e R(Reusable)*, tenham um grau satisfatório de qualidade e excelência e sejam apropriados às necessidades dos pesquisadores, exige um corpo de padrões e princípios amplamente adotados e compartilhados. Considerando este fato, propõe-se que um corpo consensual de normas e padrões consubstanciem infraestruturas que devem estar subjacentes aos processos de gestão de dados. Isto porque espera-se que as coleções de dados estejam aptas para serem utilizadas para uma grande variedade de propósitos – e não somente para as finalidades para as quais elas foram inicialmente coletadas. Para tal, elas precisam ser agregadas a outras coleções em outros sistemas, compartilhadas, acessadas, analisadas e arquivadas usando um amplo espectro de tecnologias. Essa condição torna um corpo de normas e padrões comuns infraestrutura essencial para a gestão e curadoria de dados de pesquisa. À medida que os princípios e práticas da gestão de dados de pesquisa se desenvolvem, eles começam a adquirir reconhecimento como um campo de conhecimento distinto e chamando a atenção de organizações interessadas no seu aprimoramento como, por exemplo, *DCC, Codata, GO-FAIR, DataOne, DataCite*, entre muitas outras. Nesta direção, padrões e normas comumente adotados para a gestão de dados estão tomando corpo em muitas disciplinas e setores diferentes e estão sendo redefinidos em outras disciplinas. Como resultado, práticas aprimoradas para garantir a qualidade e a durabilidade dos dados digitais estão sendo continuamente estabelecidas. (National Research Council, 2015).
- **Infraestrutura tecnológica** – Compreende um vasto conjunto de atividades, equipamentos, processos e expertises que possam viabilizar os requisitos tecnológicos operacionais necessários às ciberinfraestruturas de gestão de dados, tais como: organização lógica, física e virtual dos dados; dispositivos para processamento de alto desempenho, computação em grade e armazenamento das coleções de dados locais ou em nuvem; redes locais, comunicações, conexões externas, internet,

serviços web; aquisição/desenvolvimento de códigos científicos, *software* de *workflow*; equipamentos para análise de dados e visualização, conexões, estratégias de segurança física, lógica e de rede.

- **Infraestrutura informacional** – Compreende todo o arcabouço conceitual e teórico materializado nas práticas da ciência da informação, biblioteconomia e arquivologia, que são plenamente aplicadas à gestão de dados, como seleção, catalogação, indexação, classificação e descarte, e os instrumentos e dispositivos tecnológicos que viabilizam essas práticas, como: esquemas de representação e identificação persistentes; metadados descritivos, técnicos, administrativos, de preservação e disciplinares; tesauros, vocabulários controlados, taxonomia, ontologias, esquemas de classificação; bases de dados, repositórios e bibliotecas digitais e plataformas confiáveis para o arquivamento de longo prazo.
- **Infraestrutura de pessoal** – As inúmeras instituições de pesquisa desenvolvem os mais diversos enfoques de gestão de dados. Isto pressupõe equipes de apoio compostas por diferentes profissionais (Pinfield, Cox & Smith, 2014). Papéis como os de administrador de dados e cientistas de dados estão emergindo no mundo da ciência contemporânea e se incorporando às equipes mais tradicionais compostas por pesquisadores, técnicos de laboratório, assistentes de pesquisa e analistas; por outro lado, no âmbito das bibliotecas especializadas e dos repositórios, novos atores como bibliotecários e arquivistas de dados e curadores fazem a conexão entre a biblioteca e os laboratórios e apoiam a gestão das idiossincrasias disciplinares dos ciclos de vida dos dados (Ball, 2012). Entretanto, um requisito essencial - especialmente quando se trata dos serviços associados à curadoria - é a necessidade de conhecimento das disciplinas e domínios nos quais os dados são coletados, processados e utilizados. Sem alguma familiaridade com o problema a ser abordado, a cultura disciplinar, os objetivos a serem perseguidos, bem como com os métodos utilizados, nomenclatura e práticas dos campos em que os ativos digitais são usados, os curadores não serão capazes de tomar as decisões mais corretas para gerenciar esses ativos para uso atual e futuro (National Research Council, 2015). As equipes de gestão precisam dos papéis relacionados a seguir, que podem ser desempenhados, cada qual, por profissionais distintos ou acumuladamente - de forma mais próxima à realidade das instituições de pesquisa - por equipes menores.

    - **Pesquisadores** – personagem mais envolvido com a pesquisa e com os dados; como autor/criador/coletor dos dados/avaliador devem assegurar que os metadados disciplinares, registro dos dados (proveniência), documentação, contexto e qualidade estejam em conformidade com os padrões da comunidade/instituição.
    - **Bibliotecário de dados** – profissional de biblioteconomia com formação em gestão de dados; cataloga, indexa, organiza, apoia a publicação dos datasets; assessora o planejamento e a operacionalização dos repositórios e dos serviços de gestão de dados; apoia a curadoria por meio da construção de instrumentos de representação e padronização; idealmente conhece os fluxos de pesquisa de sua instituição; promove cursos, divulgação e material didático e assessora os pesquisadores na elaboração do plano de gestão de dados (PGD).
    - **Data steward** – [7] profissional que cuida dos dados de pesquisa de forma sustentável, por longo prazo, desde o desenho do estudo até a coleta, análise, armazenamento e compartilhamento de dados. Envolve todas as atividades necessárias para garantir que os dados de pesquisa digitais sejam localizáveis, acessíveis, interoperáveis e reutilizáveis (FAIR), incluindo gerenciamento de dados, arquivamento e reuso por terceiros.
    - **Arquivista de dados** – profissional de arquivologia, responsável pelo arquivamento e preservação de longo prazo dos dados e garantia de integridade, autenticidade e confiabilidade. Apoia o planejamento de sistemas de arquivamento confiáveis.

- **Cientista de dados** – profissional da área de ciência da computação e/ou da área disciplinar que contribui no desenvolvimento de tecnologias de análise, manipulação, visualização, modelagem, algoritmização e aplicação de metodologias avançadas, como inteligência artificial e aprendizagem de máquina.
- **Gerente de dados** – profissional da área de tecnologia da informação responsável pela manutenção e implementação de bases de dados, repositórios, sistemas de armazenamento; apoia a segurança, backups, checagem de integridade.
- **Curador de dados** – pesquisador ou profissionais de informação com conhecimento disciplinar que adiciona valor aos dados por meio de documentação, metadados, identificadores, contextualização, integração, reformatação, *mashup* etc.; promove o compartilhamento e o reuso; apoia a avaliação para a preservação e a criação de serviços.
- **Gestor – administrador de C&T** que compreende a importância dos dados no âmbito institucional, nacional e internacional; apoia a definição de políticas, negocia recursos junto às agências de fomento, implanta e-infraestruturas e adquire ferramentas, equipamentos, *software* e coleção de dados.

- **Infraestrutura organizacional** – O arcabouço infraestrutural pressupõe, assim como a governança, uma ancoragem baseada em alguma estrutura organizacional voltada para a pesquisa, como uma universidade, instituto de pesquisa, ou mesmo uma empresa, cujos empreendimentos dependem da gestão de dados. Estas organizações precisam oferecer tecnologias e ferramental, processos, políticas, recursos e treinamento para os vários e diversificados estágios da gestão de dados.

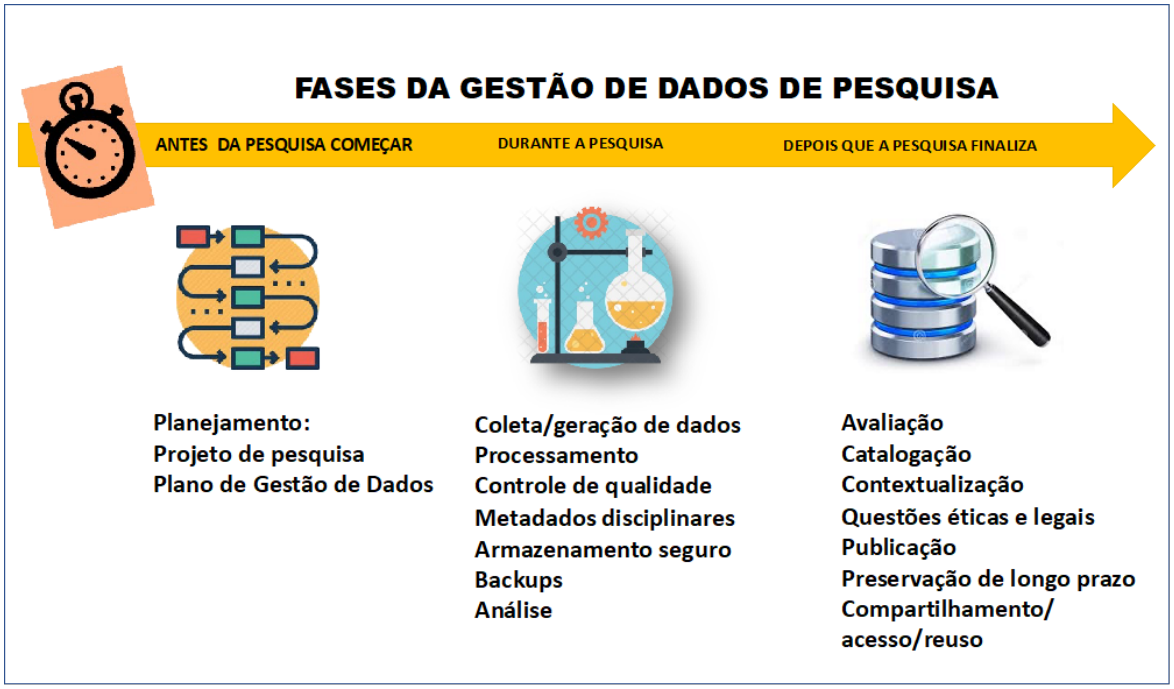
  Essas vertentes infraestruturais - que possibilitam imbricamento de saberes e práticas que estão subjacentes a equipamentos, instalações, metodologias e principalmente a pessoas - proporcionam uma vasta carteira de serviços, ferramentas e processos que continuamente levam os objetos de pesquisa para um alinhamento com os Princípios FAIR - que é um conjunto de 15 princípios estabelecidos pela comunidade de pesquisadores que quando aplicados na gestão de dados, tornam os dados FAIR.

  O que se percebe até aqui é que para se chegar na categoria que representa os serviços de gestão de dados de pesquisa, faz-se necessário o estabelecimento de uma governança de dados eficiente e a construção de uma infraestrutura adequada à formulação de serviços. Assim, o item a seguir descreve os possíveis serviços que podem ser oferecidos pelas plataformas de gestão de dados de pesquisa.

## 5. Tipos de serviços de gestão de dados

O que distingue um repositório de dados de pesquisa de uma plataforma de gestão de dados são os serviços oferecidos pela plataforma. Enquanto os repositórios têm uma função de preservação da memória da pesquisa institucional, a plataforma expande esse conceito, a partir do momento que ela inclui outros serviços. Neste sentido, o repositório passa a ser um dos serviços ofertados pela plataforma, mas que deve se integrar a outros serviços que tornem a gestão de dados mais atrativa, fácil e operacional para os usuários que farão uso dela, isto é, os pesquisadores e acadêmicos. Consideram-se então uma matriz de serviços baseados em dois eixos principais: um eixo temporal, que considera o desenrolar dos serviços de dados ao longo do tempo, interligando o ciclo de vida dos dados ao ciclo de vida da pesquisa; o segundo eixo considera o ponto de ancoragem dos serviços, significando que eles podem estar fundamentados em processos informacionais, computacionais, científicos ou administrativos. Do ponto de vista temporal, podemos considerar que a atuação da gestão na forma de serviços se efetiva em três momentos (Jones, Prior & White, 2013), conforme representada na Figura 2, a seguir, e descrita na sequência:

FIGURA 2
Fases da gestão de dados de pesquisa.

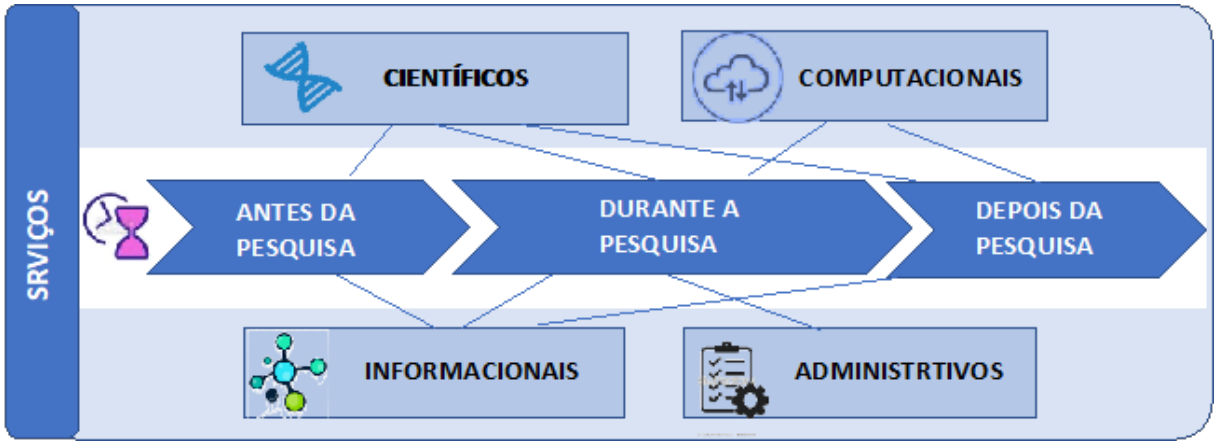



Fonte: elaboração própia.

- **Antes da pesquisa começar** – fase de planejamento e conceitualização dos dados onde é enfatizada a assistência à preparação do plano de gestão de dados, incluindo o suporte ao uso de ferramentas *on-line*, orientação sobre custos das atividades de gestão e conhecimento sobre os recursos, serviços e ferramentas disponíveis pela instituição, para a gestão de dados.
- **Durante a pesquisa** – fase que compreende um vasto conjunto de atividades informacionais, computacionais e científicas, que inclui orientação sobre documentação, formatos e padrões que potencializam a encontrabilidade, a interoperabilidade e o reuso; orientação sobre armazenamento, gerenciamento e análise de dados; aconselhamento e/ou fornecimento de sistemas de armazenamento e segurança, que atendam as necessidades de uma ampla gama de tipos de dados, plataformas e necessidades de acesso.
- **Depois que a pesquisa finalizar** – esta última fase abriga as questões de arquivamento e de acesso e preservação de longo prazo; seleção das coleções de dados de valor contínuo; suporte para tornar os dados de pesquisa disponíveis para audiências específicas; e orientação aos pesquisadores sobre como arquivar seus dados no final do projeto.

Embora esses contornos não sejam sempre bem definidos e as sobreposições estejam presentes em ambos os eixos, essa interconexão de várias expertises para a consecução das atividades de gestão de dados de pesquisa também é necessária para atender o amplo espectro das necessidades de gestão de dados. Sendo assim, conforme representado na Figura 3, a seguir, consideramos quatro tipos de serviços: serviços científicos, serviços computacionais, serviços informacionais e serviços administrativos.

FIGURA 3
Serviços de gestão de dados de pesquisa.




Fonte: elaboração própia.

## 5.1 Serviços científicos

Compreendem os serviços que se desenrolam em ambientes predominantemente científicos, como laboratórios e centros de pesquisa, e que são executados por cientistas, acadêmicos ou especialistas em gestão de dados, com profundos conhecimentos disciplinares. São serviços relacionados à preparação de dados para usos mais amplos e podem incluir atividades como, avaliação, limpeza, normalização, transformação, organização dos arquivos, nomeação e, quando necessário, anonimização e outras estratégias para a preservação da privacidade, indexação disciplinar; documentação de códigos, *workflow*, processamento e agregação de dados. Mesmo considerando que esses serviços são protagonizados pelos próprios pesquisadores, eles precisam de considerável suporte computacional e informacional e, algumas vezes, administrativo.

- **Descrição disciplinar** – para determinar se um *dataset* é realmente útil para um projeto de pesquisa, a informação descritiva (metadados e documentação) que o acompanha deve ser rica e extensiva permitindo que ele seja encontrado e interpretado, nessa direção, várias camadas de descrição podem ser aplicadas (Choudhury *et al.*, 2018). Há um consenso claro de que os pesquisadores são os profissionais mais bem posicionados para descrever os dados que eles coletam ou produzem, posto que eles compreendem com profundidade os processos pelos quais os dados foram derivados, processados, limpos, o contexto em que eles foram agregados e as limitações e fragilidades que possuem e que podem não ser aparentes para outros pesquisadores que desejam reutilizá-los no futuro (Wilson, Martinez-Uribe, Fraser & Jeffreys, 2011). Assim, o serviço deve oferecer ferramentas e instrumentos terminológicos disciplinares – metadados, ontologias, taxonomias - que permitam os pesquisadores descrever seus dados a partir do momento da sua criação (Martinez-Uribe, 2019); inclui também metodologias para o empacotamento de dados, metadados e documentação. Algumas disciplinas se utilizam do conceito de dicionário de dados que contém informação, tais como, significado dos dados, relacionamentos com outras coleções, origem, usos e formatos (Strasser, 2015). É importante salientar que uma parcela do conjunto de metadados é assinalada automaticamente pelos instrumentos científicos, exemplo: data, hora, geolocalização, temperatura etc., constituindo o conjunto de metadados intrínsecos, em contraste com os metadados assinalados pelos pesquisadores, também chamados de metadados contextuais (Mons *et al.*, 2017); os metadados descritivos (autoria, título, data de publicação etc.) são geralmente assinalados pelos profissionais de informação.

- **Avaliação** (*appraisal*) **das coleções de dados** – um dos maiores desafios em relação aos dados de pesquisa é decidir quais as coleções que precisam ser mantidas para o futuro e por quanto tempo (Martinez-Uribe, 2019). Pesquisadores ou especialistas em dados devem avaliar e selecionar as coleções que serão arquivadas por longo prazo, de acordo com orientações bem documentadas, políticas ou exigências legais. Como resultado da avaliação, alguns dados podem ser transferidos para outro custodiante ou para **destruição segura**; os dados considerados de valor contínuo são idealmente submetidos a um arquivo de dados, centro de dados, repositório ou a algum outro serviço equivalente (Ball, 2012). A avaliação também se faz necessária quando existem diversos conjuntos de dados a ser tratados e uma ordem de prioridade precisa ser estabelecida.
- **Limpeza dos dados** – consiste no processo de eliminação ou edição de parte dos dados que estão corrompidos ou sem a acurácia desejada, com o objetivo de alcançar o nível conveniente de integridade e qualidade para a coleção de dados (Sayão & Sales, 2015).
- **Organização dos dados** – consiste na organização dos dados em coleções, pastas, diretórios etc., nomeados apropriadamente, convencionados *a priori* e registrados por meio de documentos, como por exemplo, o arquivo "leia-me". Para uma organização consistente, o uso de taxonomias com princípios classificatórios bem definidos se faz necessário.
- **Transformação** – consiste na reformatação ou criação de subconjuntos de dados ou de outra derivação da coleção de dados, para reuso por pesquisadores (Ball, 2012)
- **Documentação do processamento** – os dados de pesquisa raramente são usados logo que são coletados ou gerados por um instrumento, geralmente eles passam por vários estágios de processamento que os tornam mais adequados às finalidades que se propõem. A publicação de uma descrição das etapas de processamento oferece um contexto para a interpretação e reuso dos dados e confere evidências sobre a **proveniência** dos dados (Goodman *et al*., 2014). Essa documentação pode incluir a descrição dos códigos usados na geração e processamento dos dados e do *software* de *workflow* que controlam e registram as várias etapas computacionais e de manipulação dos dados. Em algumas ocasiões, a documentação pode ser o próprio artigo onde o autor relata a metodologia de coleta e os resultados da pesquisa, ou ainda um documento textual adicional ao conjunto de dados ou um artigo de dados publicado pelo pesquisador em um periódico convencional ou periódico de dados (Torino, Roa-Martínez & Vidotti, 2020).
- **Anotação** – O reuso de dados e de outros objetos de pesquisa exige níveis elevados de colaboração. Neste sentido, os dados de pesquisa estão crescentemente localizados em plataformas que permitem novas formas de comunicação e colaboração entre acadêmicos, entre elas está a possibilidade dos pesquisadores adicionarem informações aos dados, enriquecendo-os. Este tipo de colaboração é conhecido como anotação e pode ser aplicada a todos os tipos de dados com o propósito de descrever, corrigir, interpretar, estender ou classificá-los. "Anotação é uma parte essencial da prática acadêmica [...] permitindo que o conhecimento seja organizado, compartilhado, construído e reusado" (Harvey, 2010, p. 208). O uso de vocabulários controlados ou ontologias se faz necessário para que as anotações possam ser comunicadas a outros pesquisadores e processáveis por máquina.
- **Documentação dos códigos** – O código (*software*) usado para criar ou processar os dados, em muitos casos, é um componente essencial para viabilizar o seu uso e reuso, e a documentação sobre o código é de suma importância para a compreensão dos dados e de como os resultados foram obtidos. Isso significa dizer que o código e sua documentação devem ser pensados como parte do pacote de informações que descreve os dados e, idealmente, uma cópia e uma descrição do código devem estar incluídas no pacote depositado num repositório. É preciso enfatizar que os padrões que orientam a publicação de dados estão evoluindo de forma distinta nos vários domínios disciplinares, admitindo a submissão de um amplo espectro de objetos de pesquisa; nesse contexto, os *software* desempenham vários papeis importantes no desenvolvimento de experimentos científicos, especificamente em

relação aos dados de pesquisa, porém, e em alguns casos, o *software* é o principal produto de dados, como por exemplo, na concepção de um novo algoritmo (Goodman *et al*., 2014).

- **Documentação dos** *workflows* – a combinação dos métodos de coleta dos dados, processamento e análises de um experimento é chamada de *workflow*, que minimamente indica como os dados intermediários, os produtos e os resultados finais são gerados. Em muitos casos, os pesquisadores utilizam pacotes de *software* de *workflow* para executar experimentos e registrar o que foi realizado. As informações usadas e capturadas pelo *workflow* fazem parte da **proveniência** dos dados, bem como o *software* de *workflow*, sua versão e as configurações utilizadas.
- **Análise de dados** – o serviço compreende a exploração, extração e validação de novos relacionamentos ou características de um corpo de dados (National Research Council, 2015). Na medida em que os dados são processados – transformados, normalizados, integrados - e descritos, inúmeros tipos de análises podem ser realizados sobre eles: análises quantitativas, qualitativas, visualizações, entre outras. A análise de dados é frequentemente realizada por meio de pacotes de *software* especializados, incluindo ferramentas estatísticas, *data analytics, de-identification*, processamento de linguagem natural, mineração de dados, aprendizado de máquina, algoritmos, técnicas de amostragem, desenvolvimento e teste de hipóteses. Os métodos emergentes, como a inteligência artificial, permitem análises mais refinadas, especialmente em dados não estruturados. O tipo de análise aplicada aos dados geralmente necessita de habilidades específicas, tendo como ponto de partida o conhecimento sobre o uso do pacote de *software*. Cada ferramenta de análise tem uma curva de aprendizagem, via de regra, as técnicas mais avançadas exigem conhecimentos especializados profundos (Choudhury *et al.*, 2018). Preparar os dados para a fase de análise normalmente exige conhecimentos de programação, base de dados e expertise em manipulação e edição de arquivos de dados, transformação e obtenção de saídas em diversos formatos. Estas operações que são executadas para a preparação dos dados têm que ser armazenadas e/ou descritas com o objetivo de documentar os processos e contextos e propiciar níveis apropriados de reprodutibilidade. A integridade desses processos é, naturalmente, uma preocupação e uma matéria para o pesquisador, mas que necessita de um apoio da área de computação, em termos de treinamento e assistência específica, especialmente dos cientistas de dados.
- **Apresentação e visualização de dados** – esses serviços são tipicamente compreendidos como o produto ou saída da análise de dados, são, porém, de grande importância no contexto dos serviços das plataformas de dados, posto que podem revelar uma compreensão mais acessível para um amplo espectro de interessados sobre os dados gerenciados. Envolve conhecimento sobre técnicas de visualização e apresentação, design e contextualização de informação e avaliação de produtos, algoritmos e pacotes de *software* (National Research Council, 2015).
- **Empacotamento dos dados** – para que os dados sejam encontrados por seres humanos e computadores e sejam compreendidos, agora e no futuro, é preciso que os arquivos de dados (por exemplo, planilhas) estejam fortemente acoplados a metadados e à documentação de apoio, formando uma unidade conceitual identificada chamada de pacote de dados, que pode ser preparado para ser depositado em um repositório ou centro de dados. Os exemplos variam de um único documento informacional, resultado de uma pasta *zippada*, a objetos complexos padronizados, no âmbito de uma comunidade científica, e legíveis por máquina (Tang & Hu, 2019).

Como foi mencionado, os serviços a serem ofertados podem se dividir em diversas categorias. Além dos serviços científicos aqui instanciados, os serviços computacionais constituem outra categoria de serviços necessários, que está descrita a seguir:

## 5.2 Serviços computacionais

A transição entre uma ciência fechada e autocontida para uma ciência mais aberta, distribuída em rede e cooperativa, pressupõe mudanças profundas na infraestrutura computacional necessária à condução das atividades de pesquisa, sintetizada pelo termo "ciberinfraestrutura de pesquisa". Este fato pode ser expresso pela demanda crescente de suporte computacional para a publicação de dados *FAIR*, análises integrativas avançadas, inteligência analítica (*analytics*), máquinas virtuais, sistemas de *workflow* etc. Além do mais, subjacentes aos Princípios *FAIR*, há uma ênfase especial no conceito de "acionabilidade por máquina de dados e metadados", isto reque que os recursos que desejam cumprir ao máximo as diretrizes *FAIR* devem utilizar um arcabouço tecnológico amplamente aceito, que viabilize a legibilidade por máquina de representação de dados e conhecimentos (Mons *et al.*, 2017).

Considerando esse contexto, os serviços compreendem a oferta de ferramentas de *software* e equipamentos de computação para apoiar o processamento, análise e visualização dos dados de pesquisa; apoiar os processos de interoperabilidade e acionamento por máquina de dados e metadados; prover orientação de como os dados podem melhor ser estruturados e armazenados e trabalhar, se necessário, junto aos pesquisadores na estruturação de bases de dados e marcação de texto (Wilson *et al*., 2011); os serviços podem incluir ainda treinamento específico para a equipe de pesquisadores nos recursos oferecidos e, em situações mais avançadas, oferecer processamento de alto desempenho, armazenamento em nuvem de grandes volumes e computação em grade.

- **Sistema de armazenamento** – As exigências das agências de fomento têm aumentado a conscientização dos pesquisadores quanto à necessidade de armazenar os dados de forma segura, todavia, em muitos casos, os pesquisadores armazenam em seus próprios computadores e criam sistemas informais particulares de armazenamento. Isso acontece principalmente nos estágios iniciais de um projeto de pesquisa, como durante a coleta dos dados (Choudhury *et al.*, 2018). Para minimizar esse problema, os sistemas de storage oferecem serviços críticos para a gestão de dados, colocando à disposição dispositivos de armazenamento para o amplo espectro de conjunto de dados gerados ou utilizados pelas instituições, numa escala que dê atenção aos usos correntes, mas que também antecipe os requisitos futuros das atividades de pesquisa, das diversas equipes de pesquisadores (Pinfield, Cox & Smith, 2014). Com esse horizonte, os sistemas de armazenamento performam várias ações de curto e longo prazo para garantir que os dados permaneçam seguros e íntegros, agora e no futuro, que incluem: apoio aos processos de *backups* e controles de versões; controle de acesso físico; manutenção do *hardware* de armazenamento; checagem de fixidade e atualização (*refreshing*) e migração de mídias entre outras facilidades (Ball, 2012). Os requisitos de desempenho dos sistemas de armazenamento podem variar em virtude dos níveis de utilização dos dados, por exemplo: *datasets* volumosos que serão ativamente analisados precisam de sistemas de arquivos de alta performance endereçados por diferentes camadas de armazenamento. Essas camadas de armazenamento podem ir de sistemas de arquivo paralelo de alto desempenho, que podem ser acessados por ambientes analíticos avançados, a sistemas com alta latência para acesso a *datasets* raramente usados (Choudhury *et al.*, 2018). A criação e gestão de ambientes de armazenamento exigem assistência de profissionais de computação com expertise nesse domínio técnico, isto porque a complexidade do ambiente de armazenamento local ou em nuvem pode ser "intimidante" para o pesquisador e exigir conhecimentos e ferramentas avançadas.
- **Proteção de dados sensíveis** – assegurar que os dados, especialmente aqueles classificados como confidenciais ou sensíveis, estejam mantidos seguros, anonimizados, com autenticação apropriada e mecanismos de autorização válidos (Pinfield, Cox & Smith, 2014).

- **Normalização de formatos** – nem sempre os formatos de arquivo gerados pelos instrumentos científicos são os mais adequados para a publicação, disseminação e, principalmente, para a preservação de longo prazo, posto que são formatos proprietários e muito específicos (Sayão & Sales, 2015). Assim, alguns dados podem necessitar ser migrados para formatos diferentes do original, para adequá-los às regras do sistema de arquivamento, seja para facilitar a gestão, seja para mitigar os riscos de obsolescência tecnológica, ou ambos. Nessa direção, este serviço apoia os pesquisadores quanto aos formatos, práticas e ferramentas de conversão mais adequados para produzir e documentar dados específicos; este serviço pode também oferecer suporte para projetos de bases de dados (Martinez-Uribe, 2019) e de outras formas de estruturação dos dados.
- **Serviços de** *backup* - realizar cópias de segurança é uma ação necessária para qualquer projeto de pesquisa envolvendo dados de pesquisa. A estratégia apropriada deve ser acionada pelo serviço, que pode considerar vários parâmetros: *backup* automático ou manual, testes e verificações, frequência, tempo de armazenamento, responsabilidades etc. Alguns dos processos contínuos e mais simples de backup podem ser conduzidos pelos próprios pesquisadores, mas à medida que o volume e a complexidade dos dados aumentam, eles precisam de apoio dos profissionais de computação.
- **Apoio à eliminação segura dos dados** - ao longo do processo de pesquisa, cópias de arquivos de dados que não são mais necessárias precisam ser destruídas de forma segura, principalmente as que contêm dados sensíveis. Estratégias confiáveis para apagar definitivamente arquivos de dados de pesquisa constituem um componente crítico para a gestão segura dos dados, que deve estar presente em vários estágios do ciclo de vida dos dados.

Além dos serviços científicos e computacionais, outra categoria de serviços bastante relevante é a de serviços informacionais, conforme descrita a seguir:

## 5. 3 Serviços informacionais

Grande parte dos serviços informacionais é oferecida pelas bibliotecas e executada com o apoio dos profissionais bibliotecários e arquivistas. Considerando que as bibliotecas acadêmicas historicamente têm um papel preponderante em oferecer acesso aos registros de pesquisa, nas diversas formas em que eles se apresentam, não é surpresa que a gestão de dados seja uma questão assumida globalmente pelas bibliotecas e seus profissionais (Tenopir, Birch & Allard, 2012, p. 25). Cada vez mais as bibliotecas – principalmente as que estão vinculadas às instituições de pesquisa – incorporam ao seu elenco de serviços tradicionais serviços avançados e inovadores de curadoria dos dados.

No âmbito mais amplo da gestão de dados, as responsabilidades das bibliotecas estão além dos limites de ações meramente administrativas sobre a vastidão de novos produtos de pesquisa engendrados pela ciência contemporânea. Elas podem desempenhar um papel relevante e dinâmico no desenvolvimento de esquemas de metadados, ontologias e de ferramentas que apoiem a curadoria, e em métodos de rastreamento da proveniência, no estabelecimento de políticas para o depósito e acesso a dados (Borgman, 2016, p. 13) e na reconciliação com os códigos éticos e legais vigentes. Num plano mais elevado, as bibliotecas de pesquisa podem criar estruturas de apoio à reprodutibilidade dos experimentos científicos, posto que esta noção é essencialmente baseada em registros científicos. O princípio da reprodutibilidade exige uma extensão profunda da catalogação e da indexação para incluir uma rede completa de objetos associados; requerem também uma estrutura de relacionamento de metadados elaborada que está além das práticas correntes como as dos FRBR (Functional Requirements for Bibliographic Records) [8] - além do mais, as práticas de licenciamento necessitam também se expandir para acomodar os direitos associados aos novos produtos de pesquisa. Dessa forma, os serviços informacionais compreendem um amplo espectro de atividades que vai desde o apoio à elaboração de plano de gestão de dados, até o arquivamento de longo prazo para os dados de

valor contínuo, atravessando todo o ciclo de vida dos dados, constituindo um ponto agregador e referencial de informações sobre dados. A seguir apresentamos algumas instâncias de serviços informacionais que podem ser oferecidos:

- **Portal de dados científicos –** portal *web* de dados de pesquisa de carácter institucional que tem como objetivo conectar os pesquisadores com as informações básicas sobre todos os aspectos da gestão de dados, tornando-se um ponto agregador dos serviços de gestão de dados, especialmente dos serviços de balcão de referência; o portal deve incluir também referências aos recursos externos à instituição.
- **Balcão de referência de dados –** é uma extensão do serviço de referência tradicional oferecido pelas bibliotecas de pesquisa, consistindo em um conjunto de serviços de consultoria que têm como objetivo orientar os pesquisadores nos vários aspectos da gestão de dados, tais como: identificação de repositórios para publicação; recuperação de datasets para meta-análises e outros reusos; aderência dos dados às normativas éticas, legais e de propriedade intelectual; elaboração de documento de consentimento esclarecido; atendimento aos requisitos sobre depósito de dados dos periódicos científicos; formatação de *datapapers*, entre outros.
- **Apoio na elaboração do plano de gestão de dados –** o serviço tem como objetivo assistir o pesquisador na elaboração, manutenção/revisão e consistência do plano de gestão de dados de pesquisas tendo como perspectiva o atendimento às exigências institucionais e das agências de fomento. Considera os padrões aplicáveis e ferramentas de *software* e *templates* disponíveis. O serviço pode também apoiar o desenvolvimento de padrão alinhado às diretrizes da política de dados da instituição.
- **Apoio na descoberta e acesso à coleção de dados –** descoberta de dados é o processo de encontrar e acessar dados já criados e depositados em repositórios, arquivos ou centros de dados. Neste sentido, este serviço tem como objetivo apoiar os pesquisadores na identificação, localização e acesso às coleções de dados que possam ser reusados em suas pesquisas, posto que, em muitos casos, os pesquisadores não usam somente as coleções de dados que eles criaram ou coletaram, mas também aquelas geradas e curadas por outros pesquisadores e instituições. É preciso notar que o uso de conjuntos de dados de serviços externos pode complicar o compartilhamento ou a publicação de dados posteriormente, se houver termos de licença que proíbam tais atividades.
- **Desenvolvimento de coleções de dados –** apoiar o planejamento e a construção de coleções de dados em termos informacionais, tais como: estrutura, padrões pertinentes, catalogação, metadados, identificadores, controles de versões e aderência aos padrões apropriados e às exigências éticas e legais, como por exemplo, anonimização e *copyright*. A aquisição de coleção de dados pode estar incluída neste serviço.
- **Desenvolvimento de metadados –** a aplicação de metadados confere informação e contexto aos dados, posto que isso nem sempre esteja aparente a partir deles somente, entretanto, padrões de metadados disciplinares só existem em alguns domínios, o que implica que eles precisam ser desenvolvidos (Choudhury *et al*., 2018, p. 7). Nessa direção, o serviço objetiva apoiar a estruturação e o desenvolvimento de esquemas de metadados, vocabulários, taxonomias e ontologias voltados para as especificidades disciplinares dos dados gerados pelas pesquisas desenvolvidas na instituição.
- **Criação de referências padronizadas –** assim como as publicações acadêmicas, como artigos e livros, as coleções de dados adequadamente referenciadas são mais facilmente acessadas, compartilhadas e reusadas e têm sua autoria reconhecida com mais precisão. O serviço de citação apoia o pesquisador na criação de referências padronizadas para as suas coleções de dados e de seus versionamentos de forma a aumentar a visibilidade e a citação de seus dados. A citação padronizada torna as coleções de dados e suas versões únicas e mais fáceis de ser identificadas e descobertas.

- **Identificação de dados e pesquisadores** – serviço de apoio à aquisição, criação, aplicação e manutenção de identificadores persistentes para as coleções de dados e de apoio à aplicação de identificadores a pesquisadores.
- **Catalogação/indexação das coleções de dados** – para apoiar a descoberta dos dados, metadados genéricos podem ser assinalados às coleções de dados expondo-as, dessa forma, aos grandes sistemas de descoberta, agregadores e máquinas de busca (Choudhury *et al*., 2018). Neste sentido, o serviço está voltado para a adição de metadados descritivos às coleções de dados (autor, título, identificadores persistentes etc.). Tem como referência uma política de indexação/catalogação previamente estabelecida e como objetivo aumentar os níveis de identificação e encontrabilidade dos dados. Complementa a descrição disciplinar que está focada nos métodos de coleta, processamento, análise e proveniência, e que são assinalados por pesquisadores ou especialistas de assunto. Desempenha um papel importante no registro das informações administrativas, como os termos de uso, que inclui quem pode usar os dados e como podem ser usados, incluindo também questões éticas e legais, como privacidade, tempo de embargo e problemas relacionados à propriedade intelectual.
- **Arquivamento de longo prazo/preservação** – a maioria dos dados gerenciados individualmente por pesquisadores será perdida devido à fragilidade e à degradação física das mídias nas quais os dados estão armazenados e pelo ciclo veloz da obsolescência tecnológica, porém, a perda de conhecimento ao longo do tempo se dá, principalmente, à medida que o pesquisador esquece detalhes sobre a coleção de dados e os processos de análise, e ainda por razões pessoais, como transferência, morte e aposentadoria (Mayernik *et al.*, 2012). Para mitigar esse problema, o serviço oferece infraestrutura técnica e informacional confiáveis, voltadas para o arquivamento de médio e longo prazos, de conjunto de dados selecionados e de suas informações de representação (metadados e documentação) que garantem o acesso aos seus conteúdos com níveis aceitáveis de autenticidade, confiabilidade e proveniência. O serviço tem como objetivo assegurar que as coleções de dados mantenham suas qualidades arquivísticas ao longo do tempo e do espaço, mantendo níveis de confiabilidade que viabilizam o reuso por outros pesquisadores, agora e no futuro. Para tal, é essencial empregar e se manter alinhado às normas, padrões e modelos conceituais, tais como *OAIS, RDC-arq, PREMIS*, entre outros. A expertise em curadoria assegura a resiliência e a interoperabilidade, ao longo do tempo, dos dados digitais, equacionando como os requisitos de significado, integridade, autenticidade e proveniência dos dados gerados hoje (Johnston, 2017) serão capturados e transmitidos para o futuro.
- **Publicação de dados** – a pesquisa científica nos dias de hoje não produz somente as publicações convencionais, como artigos e livros, mas produz, também, coleções de dados em vários formatos – planilhas, base de dados, modelos, algoritmos etc. Subsequentes, ou em paralelo ao atual esquema de publicação acadêmica, as coleções de dados de pesquisa podem ser publicadas de forma independente e se tornam fontes importantes de informação e de análises para novas pesquisas (Graff & Waaijers, 2011). O serviço de publicação de dados apoia a preparação de dados para a publicação, bem como assessora na escolha do repositório mais adequado em termos do tipo de dados, disciplinas, licenças, volume e eventual custo. Processos adicionais complexos estão envolvidos na preparação informacional das coleções de dados para a submissão em um repositório, incluindo a geração de metadados alinhados com esquemas relevantes e ontologias, a criação de identificadores persistentes para a gestão, não somente para a citação, mas também versões, subconjuntos e outros produtos derivativos, além dos meios mais comuns de publicação de dados, via depósito em repositórios ou centros de dados, e a publicação de uma descrição dos dados na forma de artigo de dados (*datapapers*), em periódicos especializados, chamados periódicos de dados (*datajournals*).
- **Contextualizaçãp/linking** – dependendo do campo e do projeto, uma publicação pode se basear em vários conjuntos de dados, ou um conjunto de dados pode ser base para várias publicações. Nesses

casos, os artigos se tornam descrição e documentação dos dados. Se dados e documentos relacionados pudessem ser linkados formando uma ecologia informacional, novas formas criativas de dados e informação possibilitariam pesquisas e aprendizagem distribuídas, colaborativas, multidisciplinares. Além do mais, quando alguém recuperar ou ler um artigo, pode ir diretamente aos dados, e, em alguns casos, usar as ferramentas de análise disponibilizadas pelo periódico; inversamente, quando um pesquisador acessa uma coleção de dados de seu interesse, ele pode ir diretamente aos artigos que resultaram dessas coleções. O movimento fácil entre publicações e dados [e vice-versa], nesse modelo depende de que ambos – dados e artigos – estejam acessíveis abertamente (Borgman, 2007). Dessa forma, uma infraestrutura informacional que pode estabelecer e manter conexões entre recursos associados, como dados e publicações, aumenta a cadeia de valor da pesquisa. Documentos acadêmicos e dados, nos quais eles são baseados, têm mais valor combinados, do que sozinhos.

- **Treinamento para pesquisadores** – a geração e o uso intensivo de dados da pesquisa criam novos desafios para os pesquisadores e demandam expertise em gestão de dados que, geralmente, não fazem parte da formação dos cientistas (Tenopir, Birch & Allard, 2012). Entretanto, o treinamento de usuários é essencial em todos os estágios da gestão de dados. Num ambiente em constante mudança, para que o pesquisador (público-alvo) se mantenha atualizado à medida que a infraestrutura de gestão evolui, é necessário também um estreito comprometimento dos serviços de desenvolvimento e manutenção com a capacitação dos pesquisadores (Wilson *et al*., 2011). Visando contornar esse problema, o serviço tem como objetivo dotar os pesquisadores de conhecimentos básicos sobre metadados, identificadores, plano de gestão de dados e publicação de dados. Este serviço é complementado pelo Apoio Computacional, que assiste os pesquisadores na utilização de ferramentas da tecnologia de informação, como *software* estatísticos, de visualização e *workflow*. Algumas instituições tomam como princípio o entendimento de que a capacitação em gestão de dados deve ser integrada às infraestruturas de aprendizagem acadêmica, como as disciplinas de metodologia científica.

  Soma-se ainda aos serviços científicos, computacionais e informacionais, a categoria serviços de administração, também de fundamental importância e que completa o modelo de serviços de gestão de dados de pesquisa, como observado a seguir.

## 5.4 Serviços de administração

Nesta categoria são incluídos os serviços que não se enquadram nas categoriais científicos, computacionais e informacionais, mas que são importantes para dar apoio, sustentabilidade e visibilidade àqueles serviços. Compreende serviços de orientação sobre custos, orçamento, aquisição de coleções de dados, conformidades ética e legal dos dados – especialmente dados sensíveis – às normativas e regulamentos institucionais, nacionais e internacionais; estatísticas de uso e reuso dos dados; esta categoria envolve também as questões de propriedade intelectual, licenças e tempo de embargo.

- **Aquisição de coleção de dados** – além de se utilizar de fontes externas, algumas instituições e suas bibliotecas estão adquirindo coleções individuais de dados motivadas por demandas de seus pesquisadores. Este processo muitas vezes requer uma extensa negociação em relação aos custos e à amplitude do acesso e aos termos de uso dos dados. As bibliotecas – especialmente as envolvidas com aquisição de recursos digitais – estão bem qualificadas para dar apoio a essas atividades (Choudhury *et al*., 2018).
- **Estatísticas de (re)uso** – é de grande importância para a instituição e para os pesquisadores compreender o nível de acesso e o mapa de reuso dos dados; por exemplo, que pesquisadores estão acessando os dados, em que projetos estão sendo reusados, em que áreas do conhecimento estão

sendo aplicados, que tipos de análises estão sendo realizadas. Estas informações são catalisadores importantes para a identificação de oportunidades de pesquisa colaborativa, para o aperfeiçoamento dos serviços e ainda para a prestação de contas aos financiadores da plataforma.

- **Custo/orçamento** – as atividades de gestão e compartilhamento de dados envolvem um aporte considerável de recursos financeiros em diferentes rubricas que precisam ser orçadas em um nível institucional e no âmbito dos projetos de pesquisa. Nesse contexto, as estimativas de custo, orçamentação, elaboração de cronogramas de desembolso, entre outros, se tornam uma atividade essencial, especialmente na formulação do projeto e do plano de gestão de dados.
- **Propriedade intelectual** – apoio ao pesquisador em relação às condições legais de copyright, licenças, tempo de embargo e afins, tanto para o reuso de dados de outros pesquisadores, quanto para a disponibilidade de seus próprios dados.
- **Conformidade ética e legal** – orientação sobre a conformidade ética e legal dos dados em relação à legislação nacional e internacional e aos códigos da instituição.
- **Divulgação/disseminação** - elaboração e execução de um plano de ações de divulgação/disseminação das coleções de dados que amplie as possibilidades de colaboração disciplinar e interdisciplinar e eleve o nível de visibilidade, traduzido por citações das coleções de dados.

## Conclusão

Nessa nova era científica, em que a escassez de dados e informações é menos crítica que o excesso, as dificuldades dos agentes humanos operarem, na frequência e velocidade exigidas pela complexidade das ciências intensivas de dados, reforçam a necessidade de exploradores computacionais agirem de forma autônoma e inteligente, tem-se como perspectiva a articulação de um ecossistema global de dados e serviços subjacentes aos dispositivos intelectuais, sociais e ciberestruturais de produção de conhecimento científico. Esse contexto exige mais que repositórios de dados, colocando em voga a necessidade também de outras infraestruturas que possam apoiar o desenvolvimento da pesquisa como um todo e não apenas o depósito dos dados no final da pesquisa. Neste sentido, esta pesquisa veio mostrar que além de repositórios de dados, novos serviços de gestão de dados sob plataformas mais amplas devem ser ofertados, se ajustando às infraestruturas computacionais, aos processos de análises e *workflows* sofisticados, e incorporando expertises que sejam capazes de lidar com os ambientes e processos tecnologicamente mais sofisticados da pesquisa atual.

Além disso, é preciso considerar que, na implantação de práticas e infraestruturas de gestão de dados, o contexto específico das comunidades científicas e as possibilidades da adoção devem ser observados. A importância de cada serviço proposto vai depender das prioridades e da geração e uso de determinados objetos de pesquisa. Esta condição implica que diferentes disciplinas encontrem soluções técnicas e necessitem de arcabouços infraestruturais e organizacionais em torno de serviços de gestão diferentes para alcançar o grau de FAIRificação requerido por suas comunidades.

Percebe-se ainda que para a efetiva "FAIRificação" do ecossistema de dados, a aderência não deve se aplicar somente aos dados, no sentido mais estrito, mas também aos algoritmos, ferramentas, códigos e *workflows* que levam aos dados, posto que todos os componentes dos processos de pesquisa devem estar disponíveis para assegurar a transparência, a reprodutibilidade e a reusabilidade.

Por este motivo, o presente artigo veio propor um novo conceito denominado plataforma de gestão de dados de pesquisa que visa servir como uma alternativa possível para a resolução dos diversos desafios encontrados por pesquisadores e acadêmicos, que visam encontrar, acessar, compartilhar e reusar dados como insumos para novas pesquisas. O conceito de plataforma aqui apresentado pode contribuir para que as instituições de pesquisa e financiamento estejam preparadas, não apenas para investimentos infraestruturais, mas também para o estabelecimento de políticas de incentivo pautadas em ofertas de serviços inovadores que deem suporte a todo o processo de pesquisa. A adoção deste conceito poderá ser uma solução surpreendente

para agilizar as mudanças comportamentais e organizacionais para uma ciência mais aberta, reprodutível e dinâmica.

Por fim, registramos aqui que o presente artigo é fruto de pesquisas em andamento, desenvolvidas no âmbito do grupo de pesquisa BRIET (Biblioteconomia, Representação, Interoperabilidade, *E-science* e Tecnologia), que resultam do projeto de pesquisa "Gestão de dados de pesquisa FAIR: uma proposta de modelo para aceleração das pesquisas científicas no estado do Rio de Janeiro" e do projeto de desenvolvimento "Infraestrutura de Apoio à Gestão e Preservação do Conhecimento Nuclear Brasileiro". Como resultado desses estudos, uma série de 3 artigos foram publicados, a saber: Proposta de Modelo de Serviço de Gestão de dados de Pesquisa; Modelo de Implementação para Internet de Dados e Serviços FAIR; e este sobre Plataforma de Gestão de Dados de Pesquisa.

## Agradecimentos:

Ao CNPq e à FAPERJ pelo financiamento dessa pesquisa. À Teodora Marly Gama pela revisão, sugestões e preciosa contribuição.

## Referências

Ball, A. (2012). *Review of data management lifecycle models*. Bath, UK: University of Bath. Recuperado de https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.4219&rep=rep1&type=pdf

Borgman, C. L. (2007). Data: the input and output of scholarship. In C. L. Borgman, *Scholarship in the digital age. information, infrastructure, and the Internet*. London: The MIT Press.

Borgman, C. L. (2016). *Big data, little data, no data: scholarship in the networked world*. London: The MIT Press.

Choudhury, S. *et al.* (2018). Research data curation: a framework for an institution-wide services approach. *EDUCAUSE Working Group on Data Curation, 35*. Recuperado de https://hsrc.himmelfarb.gwu.edu/libfacpubs/35

Cox, A. & Pinfield, S. (2014). Research data management and libraries: current activities and future priorities. *Journal of librarianship and information science, 46*(4), 299-316. Recuperado de http://lis.sagepub.com/cgi/doi/10.1177/0961000613492542

Fearon, D. Jr., Gunia, B., Lake, S., Pralle, B. E. & Sallans, A. L. (2013). *SPEC Kit 334: Research data management services.* Washington, DC: Association of Research Libraries.

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M. et al. (2014). Ten simple rules for the care and feeding of scientific data. *PLoS computational biology, 10*(4). Recuperado de https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003542

Graaf, M. van der & Waaijers, L. (2011). *A surfboard for riding the wave: towards a four country action programme on research data*. Copenhagen: Knowledge Exchange. Recuperado de https://www.voced.edu.au/content/ngv%3A48428

Harvey, R. (2010). *Digital curation: a how-to-do-it manual.* New York: Neal-Schuman Publisher, Inc.

Johnston, L. (2017). *Introduction to data curation from curating research data* (vol. 1). Chicago: Association of College & Research Libraries. Recuperado de https://conservancy.umn.edu/handle/11299/185334

Jones, S., Prior, G. & White, A. (2013). *How to develop research data management services – a guide for HEIs*. Edinburgh: Digital Curation Centre. Recuperado de https://www.dcc.ac.uk/guidance/how-guides/how-develop-rdm-services

Leonardi, P. M. (2010). Digital materiality? How artifacts without matter, matter. *First monday*, *15*(6-7). Recuperado de https://journals.uic.edu/ojs/index.php/fm/article/view/3036

Marín-Arraiza, P. & Vidotti, S. (2019). Implementação de serviços institucionais de administração de dados. *Liinc Em revista, 15*(2). https://doi.org/10.18617/liinc.v15i2.4819

Martinez-Uribe, L. (2019). *Research data management services: findings on the consultation with service providers.* Oxford: Oxford Digital Repositories Steering Group.

Mayernik, M. S. *et al*. (2012). The data conservancy instance: infrastructure and organizational services for research data curation. *D-Lib magazine, 18*(9-10). Recuperado de http://www.dlib.org/dlib/september12/mayernik/09mayernik.html

Mons, B. *et al.* (2017). Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud. *Information services & use, 37*(1), 49-56.

Mons, B. (2018). *Data stewardship for open science: implementing FAIR principles*. Boca Ratón: Chapman and Hall/CRC.

National Research Council. (2015). *Preparing the workforce for digital curation.* Washington, D.C.: The National Academies Press.

Nielsen, H. J. & Hjorland, B. (2014). Curation research data: the potential roles of libraries and information professionals. *Journal of documentation, 70*(2). Recuperado de https://www.emerald.com/insight/content/doi/10.1108/JD-03-2013-0034/full/html

Pinfield, S., Cox, A. M. & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PLoS one*, *9*(12), e114734. Recuperado de https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114734

Sales, L. F. & Sayão, L. F. (2018). *A ciência invisível: revelando os dados da cauda longa da pesquisa*. Em *Encontro Nacional de Pesquisa em Ciência da Informação, 19*. *Anais* ... Marília: UNESP.

Sayão, L. F. & Sales, L. F. (2015). *Guia de gestão de dados de pesquisa para pesquisadores e bibliotecários.* Rio de Janeiro: CNEN.

Sayão, L. F. & Sales, L. F. (2020). Afinal, o que é dado de pesquisa? *BIBLOS, 34*(2). Recuperado de https://www.seer.furg.br/biblos/article/view/11875

Strasser, C. (2015). *Research data management.* Baltimore: NISO. Recuperado de https://wiki.lib.sun.ac.za/images/2/24/PrimerRDM-2015-0727.pdf

Tang, R. & Hu, Z. (2019). Providing research data management (RDM) services in libraries: preparedness, roles, challenges, and training for RDM practice. *Data and information management, 3*(2), 84-102.

Tenopir, C., Birch, B. & Allard, S. (2012). *Academic libraries and research data services: Current practices and plans for the future*. Chicago, IL: Association of College and Research Libraries. Recuperado de https://trace.tennessee.edu/utk_dataone/20/

Torino, E., Roa-Martínez, S. M. & Vidotti, S. A. B. G. (2020). Dados de pesquisa: disponibilização ou publicação? Em L. F. Sales, M. Shintaku & M. Costa, *Tópicos sobre dados abertos para editores científicos*. Recuperado de http://ridi.ibict.br/handle/123456789/1072

Wilkinson, M. D. *et al.* (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data, 3*(1) 1-9. Recuperado de https://www.nature.com/articles/sdata201618.pdf

Wilson, J. A. J., Martinez-Uribe, L., Fraser, M. A. & Jeffreys, P. (2011). An institutional approach to developing research data management infrastructure. *The international journal of digital curation, 6*(2). Recuperado de http://ijdc.net/index.php/ijdc/article/view/198

## Notas

1 Recuperado de https://www.data-archive.ac.uk/
2 Recuperado de https://www.ncbi.nlm.nih.gov/genbank/
3 Recuperado de https://www.rcsb.org/
4 Recuperado de https://www.uniprot.org/
5 Recuperado de https://spdf.gsfc.nasa.gov
6 Recuperado de http://simbad.u-strasbg.fr/simbad/

7 Marín-Arraiza & Vidotti (2019) trabalharam o conceito de data steward, traduzindo, o termo como administrador de dados. De acordo com a visão das autoras, esse profissional seria o grande gestor que está na alta direção e promove a gestão de dados através da formulação de políticas. No entanto, os autores da presente pesquisa entendem que o conceito de data steward, conforme Mons (2018) que destaca que a principal habilidade deste profissional "é "ajudar os especialistas do domínio a discernir padrões significativos, correlações verdadeiras e, mais importante, a cavar as explicações mecanicistas e as relações causais que levam ao conhecimento acionável" (Mons, 2018, p.11, tradução nossa) e define data steward como "tratar os dados e os objetos de pesquisa associados com o máximo cuidado, com o objetivo de torná-los reutilizáveis para descoberta, desde que sejam válidos" (Mons, 2018, p. 24, tradução nossa).

8 Recuperado de https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr/frbr_2008.pdf