Catalogação de datasets ômicos no Repositório de Dados de Pesquisa da Embrapa

Cataloging of omics datasets in the Embrapa's Research Data Repository

Marcia Izabel Fugisawa Souza
Doutora em Educação pela Universidade Estadual de Campinas (Unicamp). Analista da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
E-mail: marcia.fugisawa@embrapa.br

Tércia Zavaglia Torres
Doutora em Educação pela Universidade Federal de São Carlos (UFSCar). Analista da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
E-mail: tercia.torres@embrapa.br

Alessandra Rodrigues da Silva
Doutora em Ciência da Informação pela Universidade de Brasília (UnB). Analista em Gestão da Informação da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
E-mail: alessandra.rodrigues@embrapa.br

Antonio Nhani Júnior
Doutor em Bioquímica pela Universidade de São Paulo (USP). Pesquisador da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
E-mail: antonio.nhani@embrapa.br

Marcos Cezar Visoli
Mestre em Informatique et Systèmes pela Université Blaise Pascal. Pesquisador da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
E-mail: marcos.visoli@embrapa.br

Carla Cristiane Osawa
Mestra em Química pela Universidade Estadual de Campinas (Unicamp). Analista da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
E-mail: carla.osawa@embrapa.br

Poliana Fernanda Giachetto
Doutora em Zootecnia pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). Pesquisadora da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
E-mail: poliana.giachetto@embrapa.br

Paula Regina Kuser Falcão
Doutora em Cristalografia de Proteínas pela University of London. Pesquisadora da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).
E-mail: paula.kuser-falcao@embrapa.br

RESUMO

O estudo teve o objetivo de definir e estabelecer regras mínimas para a catalogação de *datasets* ômicos, oriundos do Laboratório Multiusuário de Bioinformática da Embrapa (LMB), no Repositório de Dados de Pesquisa da Embrapa (Redape). Foram adotados os seguintes procedimentos metodológicos: i) revisão bibliográfica; ii) construção do estado do conhecimento sobre catalogação; iii) identificação de regras mínimas para descrição de *datasets*; iv) exploração dos elementos metadados do software Dataverse; v) identificação de atributos para descrever elementos metadados; vi) estabelecimento de regras para orientar a descrição de elementos, campos e subcampos dos esquemas Metadados de Citação e Metadados Ciência da Vida. Como resultados foram definidas e estabelecidas: i) regras mínimas de catalogação e descrição de metadados, contemplando aspectos da representação descritiva e temática; ii) instruções para o preenchimento do formulário de catalogação no Redape, uma instância do software Dataverse; iii) orientações para o uso de vocabulários controlados FAIR; iv) recomendações para revisão na catalogação de *datasets*. O artigo relata a experiência de catalogação de *datasets* ômicos no Redape, descrevendo as orientações estabelecidas para a geração de metadados qualificados, descritos de acordo com padrões internacionais de representação descritiva, e em aderência aos princípios FAIR. A aderência aos princípios FAIR assegura a interoperabilidade e o compartilhamento, bem como o uso e o reúso dos dados de pesquisa, aspectos preconizados nas diretrizes e estratégias da Política de Governança de Dados, Informação e Conhecimento da Embrapa.

Palavras–chave: Representação descritiva. Representação temática. Dados biológicos. *Datasets*. Gestão de dados de pesquisa.

ABSTRACT

The study aimed to define and establish minimum rules for the cataloging of omic datasets, coming from Embrapa's Multiuser Bioinformatics Laboratory (LMB), at Embrapa's Research Data Repository (Redape). The following methodological procedures were adopted: i) bibliographic review; ii) construction of the state of knowledge about cataloging; iii) identification of minimum rules for dataset description; iv) exploration of the metadata elements of the Dataverse software; v) identification of attributes to describe metadata elements; vi) establishment of rules to guide the description of elements, fields and subfields of the Citation Metadata and Life Science Metadata schemes. As a result, the following were defined and established: i) minimum rules for cataloging and describing metadata, covering aspects of descriptive and thematic representation; ii) instructions for completing the cataloging form in Redape, an instance of the Dataverse software; iii) guidelines for the use of FAIR controlled vocabularies; iv) recommendations for reviewing dataset cataloging. The article reports the experience of cataloging omic datasets in Redape, describing the guidelines established for the generation of qualified metadata, described according to international standards of descriptive representation, and in compliance with FAIR principles Adherence to the FAIR principles ensures interoperability and sharing, as well as the use and reuse of research data, aspects recommended in the guidelines and strategies of Embrapa's Data, Information and Knowledge Governance Policy.

Keywords: Descriptive representation. Thematic representation. Biological data. Datasets. Research data management.

1 INTRODUÇÃO

A catalogação ou representação bibliográfica é uma disciplina da Biblioteconomia e da Ciência da Informação que consiste em um conjunto de informações que simbolizam um registro do conhecimento, podendo ser definida como:

> O estudo, preparação e organização de mensagens, com base em registros do conhecimento, reais ou ciberespaciais, existentes ou passíveis de inclusão em um ou vários acervos, de forma a permitir a interseção entre as mensagens contidas nestes registros do conhecimento e as mensagens internas dos usuários. (MEY; SILVEIRA, 2009, p. 7).

A catalogação requer o emprego de formatos e padrões internacionais (de metadados[1], inclusive) e de normas técnicas para executar a representação descritiva, bem como a adoção de taxonomias, tesauros e vocabulários controlados para formular a representação temática. Isso se aplica a quaisquer artefatos, objetos e unidades de informação, sejam impressos ou digitais.

[1] Podem ser considerados como dados sobre outros dados. É o termo que os bibliotecários colocaram em catálogos e que se refere comumente à informação descritiva sobre recursos da web. Um registro de metadados consiste em um conjunto de atributos, ou elementos, necessários para descrever o recurso em questão. (MEY; SILVEIRA, 2009, p. 133).

No contexto da Ciência Aberta[2] e da Biblioteconomia de Dados[3], a catalogação de *datasets* deve ser implementada em conformidade com os princípios FAIR, para que metadados e dados sejam descritos com padrão de qualidade visando o reúso desses dados para a pesquisa. Ademais, a catalogação beneficia a transparência da pesquisa, o que favorece a descoberta de dados. Também cria condições para a reprodutibilidade e verificação da pesquisa publicada. (GORDANA; DRAGAN, 2017).

Cabe explicar que se optou pelo uso do termo *'dataset'* (no idioma inglês) por considera-lo apropriado para se referir a coleções de dados estruturados, enquanto conjuntos de dados (a forma traduzida) abarca tanto dados estruturados quanto não estruturados. "Um *dataset* no Dataverse é um contêiner dos arquivos de dados, da documentação, dos códigos e dos metadados descritivos desse *dataset.*" (DATAVERSE, 2016, tradução nossa). De acordo com Rocha et al. (2021, p. 9), no software Dataverse, "Um *dataset* é composto por metadados, pelos termos de uso (como licenças) e por arquivos."

O artigo tem o objetivo de relatar a experiência pioneira da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) em catalogação de *datasets* oriundos de pesquisas conduzidas no Laboratório Multiusuário de Bioinformática da Embrapa (LMB), conjugada a outras ações de implementação e melhoria do processo de gestão de dados científicos.

1.1 CIÊNCIA ORIENTADA A DADOS E PRINCÍPIOS FAIR

Alinhada aos pressupostos da Ciência Aberta, dos dados abertos e da *e-Science*, insere-se a ciência orientada a dados, assim denominada por ser constituída em uma ciberinfraestrutura de tecnologia de informação e comunicação e caracterizada pelo uso de plataformas abertas, aplicações, padrões de metadados, protocolos de interoperabilidade, identificadores persistentes, linguagem semântica e repositórios abertos. Ademais, a ciência orientada a dados requer o emprego de normas e padrões para

[2] "Uma nova abordagem para investigações no contexto da comunicação científica e da ciência cidadã. [...]. Está sujeita a múltiplas interpretações, sendo também denominada de "*Open Science*, *e-Science, Open Research, Research Science e Data Science.*" (OLIVEIRA; SILVA, 2016, p. 10-11). Para Albagli (2015, p. 15), "Ciência aberta passa a constituir um termo guarda-chuva, que vai além do acesso livre a publicações científicas e inclui outras frentes, como dados científicos abertos, ferramentas científicas abertas, hardware científico aberto, cadernos científicos abertos, *wikipesquisa*, ciência cidadã, educação aberta.

[3] Conceito definido por Semeler e Pinto (2019) como uma Biblioteconomia orientada a dados, derivada do conceito de ciência orientada a dados.

descrição de recursos digitais, sobretudo, dos dados de pesquisa, assim como para a citação desses dados.

De acordo com Oliveira e Silva (2016, p. 9), "O modus operandi da ciência modificou-se para acomodar os dados de pesquisa como ponto central." São inúmeras as definições de dados de pesquisa, no entanto, no contexto deste artigo, optou-se por adotar aquela que define dados de pesquisa como "registros factuais (pontuações numéricas, registros textuais, imagens e sons) usados como fontes primárias para a pesquisa científica, e que são comumente aceitos na comunidade científica como necessários para validar os resultados de pesquisa." (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2007, p. 13, tradução nossa).

Dados de pesquisa são objetos digitais e não digitais, considerados ativos altamente valiosos e imprescindíveis para o avanço do conhecimento científico. Sales et al. (2019, p. 306) definem dados de pesquisa, como:

> [...] em sua maioria, objetos digitais complexos, que carregam em si todas as idiossincrasias da pesquisa, do método, da tecnologia e da área que os originou, e cuja gestão é muito mais dependente desses parâmetros e mais específica do que a gestão generalista de produtos de pesquisa mais convencionais, como livros, artigos e relatórios.

No entanto, dados de pesquisa requerem a implementação de práticas gerenciais eficientes, que assegurem o processamento, a preservação, a integridade e a análise, bem como favoreçam o compartilhamento e o reúso, ao longo de seu ciclo de vida. (SALES et al., 2019). Dados de pesquisa, para que possam ser disponibilizados para reúso, necessitam estar organizados, documentados, descritos com padrão de metadados e catalogados em conformidade com padrões internacionais de representação descritiva, com vistas a garantir a integridade e a qualidade, em benefício da credibilidade da informação contida nesses ativos digitais. De acordo com Semeler e Pinto (2019, p. 116):

> Os dados de pesquisa precisam ser identificáveis, citáveis, visíveis, recuperáveis, interpretáveis, contextualizáveis, interoperáveis e reutilizáveis onde quesitos de consistência e procedência são considerados.

Tais exigências estão em consonância com os princípios FAIR, que se constituem de um conjunto de elementos orientadores para que diferentes *stakeholders* envolvidos no processo de gestão de dados possam aumentar a reutilização de seus acervos de dados de pesquisa. (WILKINSON et al., 2016, p. 1). Os princípios FAIR e seus elementos são apresentados no Quadro 1, abaixo.

Quadro 1 – Princípios de dados FAIR

Princípio F - *Findable* (Encontrabilidade) e seus elementos
F1. Os metadados e dados devem ter identificadores globais, persistentes e identificáveis.
F2. Os dados devem ser descritos com metadados enriquecidos (impacta diretamente R1).
F3. Os metadados devem incluir claramente e explicitamente os identificadores dos dados que descrevem.
F4. Os metadados e dados devem ser recuperáveis ou indexados em recursos que ofereçam capacidade de busca.
Princípio A - *Accessible* (Acessibilidade) e seus elementos
A1. Metadados e dados devem ser recuperáveis pelos seus identificadores usando protocolo de comunicação padronizado.
A1.1. O protocolo deve ser aberto, gratuito e universalmente implementável.
A1.2. O protocolo deve permitir procedimentos de autenticação e autorização, quando necessário.
A2. Metadados devem ser acessíveis, mesmo quando os dados não estão mais disponíveis.
Princípio I - *Interoperable* (Interoperabilidade) e seus elementos
I1. Metadados e dados devem ser representados por meio de uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento.
I2. Metadados e dados devem usar vocabulários que assegurem os *princípios* FAIR.
I3. Metadados e dados devem incluir referências qualificadas para outros metadados e dados.
Princípio R - *Reusable* (Reusabilidade) e seus elementos
R1. Metadados e dados são descritos com uma pluralidade de atributos precisos e relevantes.
R1.1. Metadados e dados devem ser disponibilizados com licenças de uso claras e acessíveis.
R1.1. Metadados e dados devem estar associados à sua proveniência.
R1.3. Metadados e dados devem estar alinhados com padrões relevantes ao seu domínio.

Fonte: Henning et al. (2019, p. 178-179).

No intuito de ampliar a compreensão desses princípios, em 2018, a *European Commission Expert Group on FAIR Data* publicou o documento *'Turning FAIR reality"*, considerado um instrumento para a aplicação dos princípios FAIR em diferentes contextos e dimensões (EUROPEAN COMMISSION EXPERT GROUP ON FAIR DATA, 2018).

Desde então, diversas iniciativas relacionadas ao ecossistema dos princípios FAIR foram criadas e outras encontram-se em desenvolvimento, orientando a sua aplicação no contexto da gestão de dados de pesquisa, para torná-los encontráveis, acessíveis, interoperáveis e reusáveis. Diante disso, o tratamento descritivo de dados e metadados, utilizando-se padrões e regras de catalogação, torna-se imprescindível para o compartilhamento e o reúso de dados em pesquisas científicas.

Entretanto, assevera Veiga (2019, p. 15), "Não basta compartilhar dados, eles precisam ser FAIR." Salienta-se, todavia, que não apenas os dados precisam ser FAIR, mas também os metadados precisam ser FAIR, pois são estes princípios que irão orientar a ampliação do acesso e possibilitar a correta interpretação e utilização dos dados pela comunidade científica.

Contudo, parte-se da premissa de que dados científicos para serem usados e reusados precisam, antes, ser encontrados, e, precipuamente, devem estar preservados. Assim, para que isso ocorra, esses dados devem estar associados a metadados FAIR. Neste particular, detectam-se pontos de interação entre princípios FAIR e a Biblioteconomia e a Ciência da Informação, em especial, no que tange às atividades de representação descritiva e temática de *datasets*. Por conseguinte, bibliotecários e cientistas da informação, em decorrência de suas formações acadêmicas, são os profissionais que possuem melhor competência para participar diretamente da atividade de catalogação de dados e de metadados. Semeler e Pinto (2019, p. 122) destacam que:

> Os dados de pesquisa são parte essencial do registro acadêmico científico e a gestão dos dados de pesquisa é cada vez mais vista como uma tarefa elementar para bibliotecas acadêmicas. Assim, constituem uma nova área de pesquisa para os chamados bibliotecários de dados.

Na literatura, a catalogação de dados (GORDANA; DRAGAN, 2017; BRANDT et al., 2019) está relacionada ao conceito de Biblioteconomia de dados definido, por Semeler e Pinto (2019), como aquela orientada a dados e associada ao conceito de ciência orientada a dados.

2 PROCEDIMENTOS METODOLÓGICOS

O estudo caracteriza-se como qualitativo do tipo revisão bibliográfica de caráter exploratório e descritivo. Optou-se pela pesquisa qualitativa do tipo revisão bibliográfica porque esta é uma abordagem que permite a contextualização do problema da catalogação de *datasets* de pesquisa no interior de uma atividade prática de pesquisa conduzida pelo LMB. Além disso, esta abordagem contribui para a construção de análises que oferecem lastro para a concepção do quadro teórico a ser utilizado na investigação empreendida (ALVES-MAZZOTTI, 2002). É uma pesquisa que também tem caráter exploratório porque observa e compreende os vários aspectos teóricos inerentes ao fenômeno estudado, a partir da exploração obtida com a revisão bibliográfica (GIL, 2017), e realiza uma descrição de como deve ser a catalogação de *datasets* no LMB (GERHARDT et al., 2009; GIL, 2017).

Os procedimentos metodológicos adotados foram: i) revisão bibliográfica; ii) compreensão do estado do conhecimento sobre catalogação e representação descritiva; iii) identificação de regras mínimas para a descrição de *datasets*; iv) identificação dos esquemas e elementos metadados do software Dataverse; v) identificação de conjuntos de atributos comuns para orientar a descrição dos elementos metadados; vi) estabelecimento de regras para orientar o preenchimento de cada elemento, campo e subcampo nos esquemas Metadados de Citação e Metadados Ciência da Vida.

3 CATALOGAÇÃO DE DADOS DE PESQUISA

O panorama atual da Ciência voltado para a produção de grandes volumes de dados, ensejado pelas possibilidades de compartilhamento e reúso, traz novos desafios para bibliotecários e cientistas da informação no tocante à gestão de dados, em especial, na catalogação descritiva de *datasets* de pesquisa.

Novas abordagens teóricas, metodológicas e técnicas vêm sendo acrescidas aos conhecimentos biblioteconômicos tradicionais para torná-los condizentes com os desafios atuais da Ciência Aberta e da ciência orientada a dados. Contudo, Semeler e Pinto (2019, p. 124) esclarecem que “A biblioteconomia orientada a dados não é um novo ramo da biblioteconomia, fundamenta-se em uma diversidade de habilidades já incorporadas e conhecidas por bibliotecários.” Nesse sentido, Zafalon (2012, p. 37) chama a atenção para

a exigência sempre atual de adoção de padrões para representação da informação em ambientes digitais, de tal sorte "[...] que tanto recuperação quanto acesso estejam garantidos pela identificação unívoca do recurso."

3.1 METADADOS E DESCRIÇÃO DE RECURSOS

Os metadados FAIR são os elementos que transportam o conteúdo das informações que devem ser expressos por meio de técnicas de representação descritiva dos dados, condição essencial para o uso e reúso de dados de pesquisa. (ROCHA et al., 2017; DIAS et al., 2019). Assim, a representação descritiva tem como grande desafio facilitar a descoberta de conhecimento, conforme assinala a comunidade FORCE11, comunidade internacional formada por pesquisadores, bibliotecários, editores e financiadores de pesquisa, dedicada ao compartilhamento de conhecimento para reformar ou aprimorar o sistema de publicação e comunicação científica (THE FUTURE OF RESEARCH COMMUNICATIONS AND E-SCHOLARSHIP, 2021a, 2021b). A adesão aos princípios FAIR é parte dos esforços para superar esse desafio.

A análise dos princípios FAIR sob a perspectiva da atividade de representação descritiva evidencia que: i) os dados devem ser descritos com metadados enriquecidos para que sejam encontráveis; ii) os metadados devem estar acessíveis mesmo quando os dados não estiverem disponíveis; iii) tanto dados quanto metadados devem ser descritos utilizando-se vocabulários controlados interoperáveis compatíveis com os princípios FAIR; iv) dados e metadados devem ser aderentes a padrões e formatos aceitos pela comunidade de usuários, como padrões de metadados, vocabulários controlados etc. para que sejam reusáveis. (THE FUTURE OF RESEARCH COMMUNICATIONS AND E-SCHOLARSHIP, 2021b, tradução nossa).

Os metadados e a catalogação de dados no escopo deste estudo são considerados na perspectiva da gestão de dados de pesquisa e do ciclo de vida dos dados. O modelo DataONE (2021), composto pelas etapas Planejar – Coletar – Assegurar – Descrever – Preservar – Descobrir – Integrar – Analisar, foi tomado como referência pois "[...] fornece uma visão geral das etapas envolvidas para o gerenciamento e preservação de dados, visando posterior uso e reutilização." (ANJOS; DIAS, 2019, p. 84).

É válido ressaltar que as etapas componentes do modelo DataONE de ciclo de vida de dados não são dependentes entre si, e não são completamente obrigatórias para sua

otimização. Essas etapas podem e devem ser realizadas de acordo com as necessidades de cada pesquisador, comunidade científica, entre outros. A atividade de descrição de metadados por meio da representação descritiva e temática de *datasets* está inserida na etapa **Descrever**, do ciclo de vida de dados, e como tal devem estar alinhadas aos princípios FAIR.

Metadados possibilitam a exploração de outras dimensões e facetas do dado, que ao serem reveladas pela catalogação passam a contribuir para a melhoria da gestão e qualidade dos dados de pesquisa da Embrapa, favorecendo a descoberta das coleções de dados para a comunidade científica. Metadados são indispensáveis para que, no futuro, os conteúdos digitais possam ser acessados e interpretados. Sem metadados, de acordo com Gray (2002), citado por Sayão e Sales (2016, slide 83), os usuários

> [...] não saberão os detalhes de como os dados foram obtidos e preparados: 1) como os instrumentos foram projetados e construídos; 2) quando, onde e como os dados foram coletados; e, 3) não terão uma descrição dos processos que levaram aos dados derivados, que são tipicamente usados para análises científicas.

Todavia, constata-se a inexistência na literatura nacional e internacional, em Biblioteconomia e Ciência da Informação, de relatos de pesquisa e de experiências relacionadas com catalogação descritiva de *datasets* de pesquisa. Ainda que o tema metadados predomine na literatura, não é perceptível uma vinculação destes com a necessidade de representá-los descritivamente.

O desenvolvimento de iniciativas que visam ao tratamento descritivo dos dados de pesquisa é imprescindível para que o compartilhamento se viabilize em condições de fornecer informações qualificadas e essenciais ao reúso desses ativos digitais. Moreira et al. (2017, p. 159) assinalaram que a descrição eficiente dos recursos informacionais e a representação adequada dos conjuntos de dados poderiam amenizar problemas como acesso parcial, superficial e de difícil compreensão dos dados e contribuiriam com a obtenção de resultados relevantes no processo de recuperação.

Metadados também são indispensáveis à interoperabilidade técnica e semântica, ou seja, sem eles os repositórios e plataformas de dados não poderão intercambiar dados e informações. Metadados são constituídos por elementos descritivos bem definidos, como por exemplo: autor, título, descrição, assunto, palavra-chave, identificador, produtor, tipo de dados, restrições de acesso, termo de uso das coleções etc., formando a partir da

catalogação de dados um corpo de informações capaz de contextualizar os dados, sobretudo, quanto à proveniência, história, natureza e propósito.

A geração de metadados enriquecidos traz benefícios diretos para a gestão de dados, com impactos positivos no arquivamento e preservação, bem como na interoperabilidade e recuperação de *datasets* de pesquisa. Dados somente serão úteis para análise se tiverem sido descritos com metadados de qualidade; para isso, recomenda-se a adoção dos princípios FAIR e o uso de normas e padrões para descrevê-los.

A descrição de metadados de *datasets* deve ser pautada na adoção de regras de catalogação para que sejam garantidas a clareza, a integridade, a precisão, a lógica e a consistência da informação (MEY, 1995, p. 7). De acordo com Silva e Silveira (2017, p. 104): "Os elementos descritivos devem ser padronizados e utilizar um conjunto comum de regras para que facilite a integração, o compartilhamento e a própria descrição dos materiais."

3.2 CATALOGAÇÃO DE DADOS NO REPOSITÓRIO DE DADOS DE PESQUISA DA EMBRAPA

A Embrapa, empresa vinculada ao Ministério da Agricultura, Pecuária e Abastecimento, do governo federal, criada em 1972 e dedicada ao desenvolvimento de pesquisa e inovação em agricultura tropical, vem acompanhando atentamente a passagem do paradigma da ciência baseado em hipótese para o paradigma de ciência orientada a dados.

Assim como inúmeras instituições científicas nacionais e internacionais, a Embrapa, em 2019, instituiu sua Política de Governança de Dados, Informação e Conhecimento (EMBRAPA, 2019). Nela foram estabelecidos princípios, diretrizes, atribuições e responsabilidades relacionados à gestão de dados, informação e conhecimento, bem como à divulgação de informações relevantes na Empresa, como mecanismos para geração, organização, tratamento, acesso, preservação, recuperação, divulgação, compartilhamento e reúso dos seus ativos de informação.

No tocante à gestão de dados de pesquisa, em 2021, a Empresa tomou a decisão estratégica de implantar o Repositório de Dados de Pesquisa da Embrapa (Redape), para abrigar o depósito e a preservação de dados de pesquisa, com vista a promover o compartilhamento e o reúso dos dados. Este repositório vem se somar a outras iniciativas já implementadas na Empresa, como é o caso do Repositório Acesso Livre à Informação

Científica da Embrapa (Alice)[4], do Repositório de Informação Tecnológica da Embrapa (Infoteca-e)[5] e do Geoinfo[6] - Infraestrutura de Dados Espaciais da Embrapa.

Dados de pesquisa produzidos pela Embrapa são ativos digitais de valor intangível, com potencial para beneficiar diretamente inúmeros pesquisadores e instituições por meio do reúso, seja pelo reaproveitamento, agregação, integração, metanálise ou reanálise (CURTY, 2019). Todavia, a implantação de repositórios de dados traz desafios de pesquisa voltados ao incremento de novas metodologias, procedimentos e práticas, relativas ao processo de gestão de dados. No que tange à catalogação de dados, o principal desafio é desenvolver diretrizes e orientações para a representação descritiva de *datasets*, para que sejam descritos e preservados adequadamente, de acordo com requisitos de qualidade para o reúso eficiente, e que atendam aos princípios FAIR, e às orientações da biblioteconomia e da catalogação de dados (SEMELER; PINTO, 2019).

Nesse contexto se insere a necessidade de implementação de melhorias de natureza metodológica, processual e técnica, especificamente no que tange ao tratamento catalográfico dos *datasets* para que os metadados sejam descritos com qualidade que atendam aos princípios FAIR, ao serem depositados no Redape. O software utilizado é o Dataverse, uma aplicação web, de código-aberto, dedicada ao compartilhamento, preservação, citação, exploração e análise de dados de pesquisa. Neste software os *datasets* são compostos por metadados descritivos, documentação, códigos e arquivos de dados. (DATAVERSE, 2021).

3.3 LABORATÓRIO MULTIUSUÁRIO DE BIOINFORMÁTICA DA EMBRAPA: A EXPERIÊNCIA DE CATALOGAÇÃO DE *DATASETS* ÔMICOS[7]

A experiência de catalogação de *datasets* foi conduzida junto ao Laboratório Multiusuário de Bioinformática da Embrapa (LMB), no escopo de um projeto-piloto integrado a uma ação gerencial de implementação de gestão de dados de pesquisa. O LMB é uma estrutura computacional de vanguarda que foi criada em 2010 nas dependências da Embrapa Agricultura Digital, localizada na cidade de Campinas, SP, para fazer frente

[4] Disponível em: https://www.alice.cnptia.embrapa.br/.
[5] Disponível em: http://www.infoteca.cnptia.embrapa.br.
[6] Disponível em: http://www.embrapa.br/geoinfo.
[7] *Datasets* gerados por tecnologias de sequenciamento de alto rendimento, que permitem extrair informações genômicas, transcriptômicas e proteômicas. (ANTONELLI et al., 2019).

aos avanços do conhecimento nas áreas de recursos genéticos, biotecnologia e melhoramento genético.

Com a missão de viabilizar soluções de Bioinformática para projetos de pesquisa, desenvolvimento e inovação na Embrapa em um ambiente colaborativo, o LMB compartilha equipamentos de alto custo, otimiza recursos econômicos e humanos, incorpora e disponibiliza competências na área de computação de alto desempenho, para a Embrapa e para qualquer outra instituição de ensino, pesquisa, desenvolvimento e inovação do Brasil e do mundo relacionada à agropecuária. (EMBRAPA INFORMÁTICA AGROPECUÁRIA, 2021).

No LMB a implementação de processos de gestão de dados é de fundamental importância porque otimiza o processamento de alto desempenho, facilitando a execução de análises que visam à descoberta, bem como, favorecem o compartilhamento e o reúso dos dados. Isso implica dizer que a gestão de dados ômicos e, consequentemente, o processo de catalogação de *datasets*, além de serem processos contíguos e inerentes à missão do LMB, também são determinantes para assegurar o compartilhamento, favorecer o reúso e garantir a sustentabilidade dos dados.

Dados ômicos são gerados pelas tecnologias de sequenciamento de alto rendimento, que permitem extrair informações genômicas, transcriptômicas e proteômicas em grande escala de maneira rápida, confiável e barata, sendo também conhecidos como dados multi-ômicos. Alguns tipos de dados ômicos que podem ser extraídos da: 1) Genômica: i) polimorfismos de nucleotídeo único; ii) variantes raras; iii) variação no número de cópias; 2) Epigenômica: i) metilação de DNA; 3) Transcriptômica: i) expressão de RNA mensageiro (mRNA); ii) expressão de micro RNA (miRNA); iii) expressão de RNA longo não codificante (lncRNA); 4) Proteômica: i) expressão de proteína. (ANTONELLI et al., 2019, p. 267-268, tradução nossa).

O processo de catalogação de *datasets* se desenvolveu no ambiente do software Dataverse, utilizando-se um formulário para descrição dos metadados e upload de arquivos de dados. O software Dataverse oferece ao usuário a possibilidade de escolher qual conjunto de metadados deseja usar para melhor atender aos interesses e ao contexto dos dados a serem descritos.

O esquema de metadados *Citation Metadata*, doravante denominado Metadados de Citação, é predefinido pelo software Dataverse, sendo exportável para padrões internacionais, como o Dublin Core. Por conseguinte, este é o esquema de metadados

adotado no LMB para a catalogação de *datasets.* No entanto, para atender diretamente ao domínio dos dados biológicos, o Dataverse oferece também o esquema de metadados específicos - o *Life Science Metadata* -, aqui denominado de Metadados Ciência da Vida, julgado mais apropriado para descrever *datasets* ômicos.

Desse modo, os esquemas Metadados de Citação e Metadados Ciência da Vida podem ser utilizados concomitantemente, recomendação que os estudos do LMB perceberam em relação aos dados ômicos. A utilização simultânea dos dois esquemas está condicionada à estruturação do *dataverse*[8] ao qual o *dataset* estiver vinculado, ou seja, o dataverse quando criado deve prever o uso de ambos os esquemas.

O processo de catalogação de *datasets* inicia-se com a atividade de **autodepósito**, a qual deve ser executada pelo autor do *dataset* ou outra pessoa sob sua responsabilidade. O momento do autodepósito do *dataset* no software Dataverse corresponde à fase de **pré-catalogação**, quando são preenchidos os campos correspondente aos elementos metadados requeridos, que são: título autor, contato, descrição, assunto e depositante.

Na fase seguinte ocorre a **catalogação** propriamente dita, começando pela revisão do preenchimento prévio dos conteúdos relativos aos elementos metadados requeridos. A descrição catalográfica de *dataset* se ocupa essencialmente do preenchimento dos elementos metadados recomendados, fortemente recomendados e opcionais, a ser executado pelo autor do *dataset*, na função de catalogador, exigindo deste um conhecimento mínimo de normas e padrões oriundos da Biblioteconomia e da Ciência da Informação.

O emprego de técnicas de catalogação para descrever e representar dados ou quaisquer objetos digitais de pesquisa, tomando por base um conjunto estruturado de elementos metadados FAIR, é essencial para garantir a interoperabilidade técnica e semântica, o compartilhamento, o uso e o reúso dos dados de pesquisa. Ao bibliotecário e/ou cientista da informação cabe a responsabilidade técnica por esta atividade.

4 REGRAS GERAIS PARA CATALOGAÇÃO DE *DATASETS* ÔMICOS NO REPOSITÓRIO DE DADOS DE PESQUISA DA EMBRAPA

As orientações para o preenchimento do formulário de catalogação no Redape, baseadas nas regras mínimas da AACR2 (2ª edição do Código de Catalogação Anglo-

[8] *Dataverse* é aqui entendido como um projeto que reúne conjuntos de dados *(datasets).*

Americano), visam à clareza e à padronização das informações. Todos os elementos metadados, campos e subcampos, exigem atenção no preenchimento a fim de evitar erros de qualquer natureza. O preenchimento correto dos campos e subcampos tem impacto direto sobre a qualidade dos *datasets* descritos.

Dessa forma, mecanismos de busca tanto do software Dataverse como de outros buscadores na internet, sobretudo de repositórios de dados de pesquisa, tendem a recuperar os dados com mais facilidade.

As regras básicas estabelecidas que orientam o preenchimento do formulário de catalogação consistem de: i) Fornecimento de elementos metadados requeridos, fortemente recomendados, recomendados e opcionais; ii) uso de pontuação; iii) uso de maiúscula; iv) não utilização da tecla enter; v) uso de ajuda rápida; controle e padronização do nome do autor e nome corporativo; vi) controle e padronização do e-mail pessoal e corporativo; vii) controle e padronização da afiliação; viii) uso de palavras-chave e termos livres; ix) uso de padrão para descrição de datas.

5 DESCRIÇÃO DOS ELEMENTOS DO ESQUEMA METADADOS DE CITAÇÃO

Para a catalogação de *datasets*, o software Dataverse possui um esquema de metadados predefinidos denominados de Metadados de Citação. Esse esquema de metadados é formado por 33 elementos, compostos por campos e subcampos

O Quadro 2 apresenta os 33 elementos do esquema Metadados de Citação, excetuando-se campos e subcampos, que compõem o formulário de catalogação no software Dataverse. Os elementos metadados listados no idioma português foram traduzidos pelos autores, a partir da forma original em inglês, e que são mostrados entre parênteses.

Quadro 2 – Conjunto de elementos metadados do esquema Metadados de Citação do formulário de catalogação de *datasets*, nos idiomas português e inglês (língua original)

Elemento metadado	**Exigência**
Título *(Title)*	Requerido
URL alternativa *(Alternative URL)*	Recomendado
Outro identificador *(Other ID)*	Opcional
Autor *(Author)*	Requerido
Contato *(Contact)*	Requerido
Descrição *(Description)*	Requerido
Assunto *(Subject)*	Requerido

Palavras-chave *(Keyword)*	Fortemente recomendado
Classificação *(Topic classification)*	Fortemente recomendado
Publicação relacionada *(Related publication)*	Opcional
Notas *(Notes)*	Opcional
Idioma *(Language)*	Opcional
Produtor *(Producer)*	Recomendado
Data de produção *(Production date)*	Fortemente recomendado
Local de produção *(Production place)*	Recomendado
Colaborador *(Contributor)*	Fortemente recomendado
Financiamento *(Grant information)*	Recomendado
Distribuidor *(Distributor)*	Fortemente recomendado
Data de distribuição *(Distribution date)*	Recomendado
Depositante *(Depositor)*	Requerido
Data de depósito *(Deposit date)*	Recomendado
Período de cobertura de dados *(Time period covered)*	Recomendado
Período de coleta de dados *(Date of collection)*	Recomendado
Tipos de dados *(Kind of data)*	Recomendado
Série *(Series)*	Opcional
Software *(Software)*	Recomendado
Material selecionado *(Related material)*	Opcional
Datasets relacionados *(Related datasets)*	Opcional
Outras referências *(Other references)*	Opcional
Fontes de dados *(Data sources)*	Opcional
Origem das fontes de dados *(Origin of sources)*	Opcional
Características das fontes de dados observadas *(Characteristic of sources noted)*	Opcional
Documentação e acesso a fontes de dados *(Documentation and access to sources)*	Opcional

Fonte: Souza et al. (2020).

A título de demonstração, a seguir, são apresentados os principais elementos metadados requeridos e fortemente recomendados com a respectiva orientação para a descrição do *dataset*, no formulário de catalogação dos elementos do esquema Metadados de Citação (Quadro 3).

O conjunto completo de elementos do esquema (campos e subcampos) e as orientações para catalogação de *datasets* estão reunidos no documento intitulado "Catalogação de dataset no Repositório de Dados da Embrapa: a experiência do projeto-piloto de implantação de gestão de dados de pesquisa no Laboratório Multiusuário de Bioinformática da Embrapa" (SOUZA et al., 2020).

Quadro 3 - Elementos metadados, campos e subcampos requeridos e fortemente recomendados do esquema Metadados de Citação do formulário de catalogação de *datasets*

Título*

Identificador	Título
Ocorrência	Limitada
Exigência de preenchimento	Requerido
Definição	Nome atribuído ao *dataset*, e pelo qual é formalmente conhecido. Título deve ser formado pela estrutura a seguir: **Tipo de estudo ômico + Amostra estudada + Espécie estudada (nome comum + nome científico)**.
Comentário	Descrever o título do *dataset* observando as normas da língua culta, quanto à grafia e acentuação. Iniciais de título, de nome próprio e de nome científico devem ser grafadas em maiúscula.
Exemplo	Transcriptoma da glândula salivar do carrapato bovino *(Rhipicephalus (B.) microplus).*
Documento orientador	RIBEIRO, Antonia Motta de Castro Memória. **Catalogação de recursos bibliográficos pela AACR2 2002:** Anglo-American Cataloguing Rules, 2nd edition, 2002 revision. Brasília, DF: Ed. do Autor, 2003. 1 v. Paginação irregular.

Nome do autor*

Identificador	Nome
Ocorrência	Ilimitada
Exigência de preenchimento	Requerido
Definição	Nome(s) do(s) autor(es) do *dataset*. Pode(m) ser autor(es) pessoal(is) ou autor corporativo.
Comentário	Quando o autor é pessoal, preenche-se com o(s) nome(s) de pessoa(s), no formato sobrenome(s) (primeiro nome da família), seguidos dos prenomes. Caso seja um autor corporativo, descreve-se o nome da instituição, sem abreviações.
Exemplo	Nomes de autores pessoais: Silva, José Pereira Júnior, João Nome de autor corporativo: Embrapa (nome síntese) Empresa Brasileira de Pesquisa Agropecuária
Documento orientador	RIBEIRO, Antonia Motta de Castro Memória**. Catalogação de recursos bibliográficos pela AACR2 2002:** Anglo-American Cataloguing Rules, 2nd edition, 2002 revision. Brasília, DF: Ed. do Autor, 2003. 1 v. Paginação irregular.

Contato

Identificador	Contato
Ocorrência	Ilimitada
Exigência de preenchimento	Requerido

Definição	Nome da(s) pessoa(s) ou da instituição para contato sobre o *dataset.*
Comentário	É composto por três subcampos: Nome, Afiliação e E-mail.
Exemplo	Não se aplica.
Documento orientador	Não se aplica.

E-mail do contato*

Identificador	E-mail
Ocorrência	Ilimitada
Exigência de preenchimento	Requerido
Definição	Endereço de e-mail do contato sobre o *dataset*, seja pessoal ou corporativo.
Comentário	Preencher com o e-mail do contato.
Exemplo	jose.silva@empresa.com.br
Documento orientador	Não se aplica.

Descrição

Identificador	Descrição
Ocorrência	Ilimitada
Exigência de preenchimento	Requerido
Definição	Um sumário descritivo e sucinto contendo o propósito, a natureza e o escopo do *dataset*.
Comentário	É composto por dois subcampos: Texto e Data
Exemplo	Não se aplica
Documento orientador	Não se aplica.

Assunto*

Identificador	Assunto
Ocorrência	Limitada
Exigência de preenchimento	Requerido
Definição	Áreas de domínio específicas e relevantes para o *dataset.*
Comentário	Selecionar a opção de área de domínio com as quais o *dataset* está relacionado. Pode-se selecionar mais de uma opção.
Exemplos	*Agricultural Sciences* *Medicine, Health and Life Sciences*
Documento orientador	Não se aplica.

Palavras-chave

Identificador	Palavras-chave
Ocorrência	Ilimitada
Exigência de preenchimento	Fortemente recomendado
Definição	Palavras que expressam o conteúdo-chave do *dataset*, sendo capazes de representar aspectos relevantes para a

	compreensão do assunto. Palavras-chave podem ser utilizadas na construção de índices, na realização de buscas e na recuperação de informação. Palavras-chaves podem estar vinculadas ou não a um vocabulário controlado.
Comentário	É composto de três subcampos: Termo-chave, Vocabulário e *URL* do termo-chave.
Exemplo	Não se aplica.
Documentos orientadores	Não se aplica.

Termo-chave

Identificador	Termo-chave
Ocorrência	Ilimitada
Exigência de preenchimento	Fortemente recomendado
Definição	Termos-chave são indicados para representar aspectos importantes de um *dataset*. Podem ser usados na construção de índices, realização de busca e recuperação de informação. Recomenda-se a utilização de vocabulário controlado.
Comentário	Preencher com termos-chave que descrevem aspectos relevantes do *dataset.* Podem estar vinculados a um vocabulário controlado ou não.
Exemplos	Host-parasite relationships Nucleotide sequences *Rhipicephalus (Boophilus) microoplus* Ticks Transcriptome Carrapato
Documentos orientadores	*NAL Thesaurus* *NCBI Taxonomy* *AGROVOC Thesaurus*

Vocabulário controlado do termo-chave

Identificador	Vocabulário controlado do termo-chave
Ocorrência	Ilimitada
Exigência de preenchimento	Fortemente recomendado
Definição	Vocabulário controlado utilizado para a atribuição de termos-chave do *dataset*.
Comentário	Preencher com nome do vocabulário controlado utilizado para a atribuição de termos-chave para o *dataset*.
Exemplos	*NAL Thesaurus* *NCBI Taxonomy* *AGROVOC Thesaurus*
Documento orientador	Não se aplica.

URL do vocabulário controlado

Identificador	URL do termo-chave
Ocorrência	Ilimitada
Exigência de preenchimento	Fortemente recomendado

Definição	URL do vocabulário controlado onde o termo poderá ser acessado.
Comentário	**Atenção:** Neste campo não utilizar a URL do vocabulário controlado; em seu lugar preencher o campo com a URL válida do termo-chave (e não a URL do vocabulário controlado). **Observação:** No formulário Dataverse não há campo destinado à descrição da **URL do termo-chave**, por esta razão, escolheu-se o campo **URL do vocabulário controlado** para fazer o seu registro.
Exemplos	URL do termo-chave 'nucleotide sequence' *(AGROVOC Thesaurus)*: http://aims.fao.org/aos/agrovoc/c_27583 URL do termo-chave 'nucleotide sequences' *(NAL Thesaurus):* https://agclass.nal.usda.gov/mtwdk.exe?k=default&l=60&w=15979&s=5&t=2
Documento orientador	Não se aplica.

Classificação

Identificador	Classificação
Ocorrência	Ilimitada
Exigência de preenchimento	Fortemente recomendado
Definição	Refere-se à classificação do assunto do *dataset*, por meio da atribuição de uma ou mais categorias, sob as quais podem ser agrupados *dataset* de natureza semelhante. Recomenda-se o uso de categorias vinculadas a vocabulário controlado.
Comentário	É composto de três subcampos: Termo de classificação, Vocabulário controlado da classificação e URL do termo de classificação
Exemplo	Não se aplica.
Documento orientador	Não se aplica.

Termo de classificação

Identificador	Termo de classificação
Ocorrência	Ilimitada
Exigência de preenchimento	Fortemente recomendado
Definição	Indicação da classificação de assunto (categoria) de maior abrangência, na qual o *dataset* se classifica.
Comentário	Preencher com a expressão que indique a classificação de assunto mais abrangente para representar o *dataset*. O uso de termos associados a vocabulários controlados é recomendado.
Exemplos	Animal Science and Animal Products *(NAL Thesaurus)* Phenomena *(AGROVOC Thesaurus)* Ixodida *(NCBI Taxonomy)*
Documento orientador	Não se aplica.

Vocabulário controlado da classificação

Identificador	Vocabulário controlado da classificação
Ocorrência	Ilimitada
Exigência de preenchimento	Fortemente recomendado
Definição	Vocabulário controlado utilizado para a classificação de assunto do *dataset*.
Comentário	Preencher com nome do vocabulário controlado utilizado para a atribuição da classificação de assunto para representar o *dataset.*
Exemplos	*NAL Thesaurus* *NCBI Taxonomy* *AGROVOC Thesaurus*
Documento orientador	Não se aplica.

URL do vocabulário controlado

Identificador	URL do termo de classificação
Ocorrência	Ilimitada
Exigência de preenchimento	Fortemente recomendado
Definição	URL do vocabulário controlado onde o termo de classificação de assunto (categoria) do *dataset* poderá ser acessado.
Comentário	**Atenção:** Neste campo não utilizar a URL do vocabulário controlado; em seu lugar preencher o campo com a URL válida do termo-chave (e não a URL do vocabulário controlado). **Observação:** No formulário Dataverse não há campo destinado à descrição da **URL do termo-chave**, por esta razão, escolheu-se o campo **URL do vocabulário controlado** para fazer o seu registro.
Exemplos	URL do termo de classificação 'phenomena' *(AGROVOC Thesaurus)*: http://aims.fao.org/aos/agrovoc/c_330704 URL do termo de classificação *'L Animal Science and Animal Products' (NAL Thesaurus):* https://agclass.nal.usda.gov/mtwdk.exe?k=default&l=60&s=cid&w=100 URL do termo de classificação 'Ixodida' *(NCBI Taxonomy)*: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=6935&lvl=3&p=nuccore&lin=f&keep=1&srchmode=1&unlock
Documento orientador	Não se aplica.

Fonte: Souza et al. (2020).

6 DESCRIÇÃO DOS ELEMENTOS DO ESQUEMA METADADOS CIÊNCIA DA VIDA

O esquema Metadados Ciência da Vida abarca as especificidades e os interesses da comunidade de dados biológicos, tornando-o apropriado para descrever *datasets* ômicos.

O Quadro 4 apresenta os nove elementos, cujo preenchimento é opcional, do esquema Metadados Ciência da Vida, que compõem o formulário de catalogação no software Dataverse. Os elementos metadados listados no idioma português foram traduzidos pelos autores, a partir da forma original em inglês, e que são mostrados entre parênteses.

Quadro 4 – Elementos do esquema Metadados Ciência da Vida do formulário de catalogação

Elemento metadado	**Exigência**
Tipo de design *(Design type)*	Opcional
Tipo de fator *(Factor type)*	Opcional
Organismo *(Organism)*	Opcional
Outro organismo *(Other organism)*	Opcional
Tipo de medida (Measurement type)	Opcional
Outro tipo de medida *(Other Measurement type)*	Opcional
Tipo de tecnologia *(Technology type)*	Opcional
Plataforma tecnológica *(Technology platform)*	Opcional
Tipo de célula *(Cell type)*	Opcional

Fonte: Souza et al. (2020).

7 VOCABULÁRIOS CONTROLADOS

Vocabulário controlado é comumente definido como uma lista autorizada de assunto, um vocabulário fechado, um tesauro. Cunha e Cavalcanti assim definem vocabulários controlados:

> Conjunto de termos que, nos sistemas de informação, devem ser empregados tanto no momento da indexação como no da recuperação. A finalidade principal desse controle é fazer coincidir a linguagem do pesquisador com a do indexador. Nos vocabulários controlados são feitas remissivas dos sinônimos e quase-sinônimos para o termo selecionado como descritor. (VOCABULÁRIO..., 2008, p. 378).

Um dos objetivos de um vocabulário controlado é compatibilizar a linguagem utilizada por pesquisadores à linguagem adotada em sistemas de informação. Em processos de catalogação, os vocabulários controlados são utilizados para representar, por meio de termos ou palavras-chave, o conteúdo do recurso de informação que está sendo catalogado. A adoção de um vocabulário controlado para descrever e classificar assuntos é uma medida necessária para prover serviços de busca e recuperação de dados e informações com mais garantia de qualidade.

Vocabulários controlados são altamente recomendados para a representação temática de *datasets*, sobretudo quando são compatíveis com os princípios FAIR,

especialmente com o princípio *Interoperable* (Interoperabilidade) (HENNING et al., 2019; VEIGA, 2019). Metadados bem descritos com vocabulários controlados e indexados em catálogos e repositórios de dados são facilmente 'encontráveis' por mecanismos de busca. Quando isso acontece, tanto dados como metadados estarão ao alcance dos usuários.

Na mesma direção, metadados descritos com vocabulários controlados permitem a representação de assuntos e de classificação de forma padronizada, garantindo, assim, a acessibilidade e a interoperabilidade semântica e legível por máquina (VEIGA, 2019).

No Redape o uso de vocabulários controlados tem as seguintes funções: i) atribuir palavras-chave e classificação de assuntos na catalogação de *datasets*; ii) indicar relacionamentos e associações hierárquicas e de equivalência entre termos atribuídos pelo catalogador de *datasets* a palavras-chave e classificação; iii) eliminar polissemia e ambiguidade na atribuição de termos nos campos Palavras-chave e Classificação; iv) possibilitar a identificação e recuperação de *datasets* descritos sob determinados termos; v) explorar a riqueza terminológica existente nos diferentes vocabulários, sob a perspectiva da análise semântica e de mineração de textos; vi) 0ferecer possibilidade de análise conceitual dos termos em diferentes idiomas, subsidiando a descrição de assunto e de classificação de *datasets*.

No formulário de catalogação do software Dataverse é possível fazer o uso de termos de vocabulário controlado em **Palavras-chave** e também em **Classificação**. Em **Palavras-chave**, recomenda-se usar o termo mais específico possível para classificar o *dataset*. Em **Classificação**, recomenda-se utilizar termos mais abrangentes.

Para a catalogação de *datasets* ômicos no Redape foram elaboradas as orientações para o uso dos vocabulários controlados *AGROVOC Thesaurus* (FAO, 2020), *NAL Thesaurus and Glossary* (ESTADOS UNIDOS, 2020) e *NCBI Taxonomy* (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2021), identificados junto à equipe do projeto-piloto como apropriados à catalogação de *datasets* no LMB. Esses vocabulários controlados são aderentes aos princípios FAIR, ou seja, os termos do *AGROVOC*, do *NAL Thesaurus* e do *NCBI Taxonomy* são encontráveis, interoperáveis, acessíveis e reusáveis.

8 REVISÃO NA CATALOGAÇÃO DE *DATASETS* ÔMICOS

A busca de qualidade na catalogação de *datasets* deve ser assegurada por medidas de controle voltadas ao atendimento a padrões e regras mínimas de catalogação, no

tocante ao preenchimento de elementos, campos e subcampos do formulário de catalogação no software Dataverse. Para tanto, duas medidas são necessárias: i) observância ao documento orientador de catalogação de *datasets* no Redape (SOUZA et al., 2020); ii) estabelecimento dos elementos de controle de qualidade na catalogação, baseado em Branco et al. (2014).

Para o controle de qualidade foram estabelecidos os seguintes elementos: i) completeza das informações registradas sobre o *dataset;* ii) precisão da descrição quanto: ortografia, pontuação, descrição textual, transcrições de *URI, URL, DOI, ORCID* etc; iii) clareza e integridade das informações registradas sobre o *dataset*; iv) fidelidade aos padrões e normas adotados para a catalogação de *datasets*; v) confiabilidade das informações registradas sobre o *dataset*; vi) coerência e consistência dos dados e informações registradas sobre o *dataset.*

Os apontamentos resultantes da atividade de controle de qualidade da catalogação reforçam a necessidade de adoção dos padrões e orientações técnicas, prescritos para a catalogação de dados ômicos no Redape, com o fim de obter níveis aceitáveis de qualidade das informações descritas, contidas nos metadados.

Como pode ser inferido pela leitura do Quadro 8, faz-se relevante enfatizar a necessidade da adoção dos padrões e orientações técnicas prescritos para a catalogação de dados ômicos no Redape, com o fim de obter níveis aceitáveis de qualidade das informações descritas, contidas nos metadados.

Além disso, a revisão da catalogação de *datasets* ômicos no Redape deve ser realizada a partir de uma matriz de atividades, que apresenta orientações para o registro de erros e/ou inconsistências, que devem ser objeto de correção. Nela é possível fazer o registro das informações que devem ser revisadas na catalogação, como: Elemento (campo/subcampo); Descrição do erro e/ou inconsistência; Orientação para correção; Elemento(s) de impacto na qualidade; Documento orientador.

9 CONSIDERAÇÕES FINAIS

O momento atual de rápida expansão dos repositórios de dados de pesquisa em todo o mundo apresenta novos desafios metodológicos, tecnológicos e técnicos, sobretudo, aos bibliotecários e cientistas da informação. Torna-se evidente a necessidade desses profissionais aumentarem suas atuações e envolvimento com a Ciência Aberta e *e-Science*,

a ciência orientada a dados, a gestão de dados de pesquisa e áreas correlatas, a fim de melhor compreenderem as inter-relações entre esses novos domínios de conhecimento e a Biblioteconomia e a Ciência da Informação. O estudo aqui relatado se insere nesse contexto.

O estudo conduzido no LMB em parceria com o Grupo de Pesquisa em Engenharia da Informação teve o objetivo de produzir orientações para catalogação de *datasets* ômicos no Redape. Essas orientações são compatíveis com padrões internacionais e com as regras mínimas de representação descritiva. Apoiadas nos princípios FAIR, tais orientações possibilitam que: i) os dados sejam descritos com metadados enriquecidos para que sejam encontráveis; ii) os metadados descritos possam se tornar acessíveis mesmo quando os dados não estiverem mais disponíveis; iii) os dados e os metadados descritos com vocabulários controlados interoperáveis tornam-se compatíveis com os princípios FAIR; iv) os dados e os metadados descritos com padrão de metadados e vocabulários controlados aceitos pela comunidade de bioinformática possam ser reusáveis.

Ademais, trata-se de uma importante contribuição para a implementação da Política de Governança de Dados, Informação e Conhecimento da Embrapa, visto que a catalogação é uma atividade essencial para a interoperabilidade, o compartilhamento, o uso e o reúso dos dados de pesquisa, aspectos preconizados nas diretrizes e estratégias desta política.

REFERÊNCIAS

ALBAGLI, Sarita. Ciência aberta em questão. In: ALBAGLI, Sarita; MACIEL, Maria Lucia; ABDO, Alexandre Hannud. (org.). **Ciência aberta, questões abertas**. Brasília, DF: Ibict; Rio de Janeiro: Unirio, 2015. cap. 1, p. 9-25.

ALBAGLI, Sarita; APPEL, Andre Luiz; MACIEL, Maria Lucia. *e-Science* e ciência aberta: questões em debate. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais [...].** Florianópolis: ANCIB, 2013. 19 p. Disponível em: https://ridi.ibict.br/bitstream/123456789/465/1/Sarita2.pdf. Acesso em: 16 ago. 2021.

ALVES-MAZZOTTI, A. J. A "revisão bibliográfica" em teses e dissertações: meus tipos inesquecíveis – o retorno. In: BIANCHETTI, L.; MACHADO, A. M. N. (org.). **A bússola do escrever:** desafios e estratégias na orientação de teses e dissertações. Florianópolis: Editora da UFSC, 2002. p. 25-44.

ANJOS, Renata Lemos dos; DIAS, Guilherme Ataíde. A. Atuação dos profissionais da informação no ciclo de vida dos dados – DataONE: um estudo comparado. **Informação & Informação**,

Londrina, v. 24, n. 1, p. 80–101, jan./abr. 2019. Disponível em: http://eprints.rclis.org/34342/1/dataone_paper.pdf. Acesso em: 15 ago. 2021.

ANTONELLI, Laura; GUARRACINO, Mario Rosário; MADDALENA, Lucia; SANGIOVANNI, Mara. Integrating imaging and omics data: a review. **Biomedical Signal Processing and Control**, Oxford, v. 52, p. 264-280, July, 2019. DOI: https://doi.org/10.1016/j.bspc.2019.04.032.

BRANCO, Zuleica de Souza; MACHADO, Denise Ramires; CESTARI, Beatriz Helena Pires de Souza; OLIVEIRA, Zita Prates. Controle de qualidade em catalogação cooperativa. *In*: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 18., 2014, Belo Horizonte. **Anais [...]**. Belo Horizonte: UFMG, 2014. 21 p.

BRANDT, Mariana Baptista; VIDOTTI, Silvana Aparecida Borsetti Gregório; SANTOS, Plácida Leopoldina Ventura Amorim da Costa; ZANALON, Zaira Regina. Catalogação de metadados de negócio a partir dos princípios e objetivos bibliográficos. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 23, n. 4, p. 3-18, jul./set. 2019. DOI: https://doi.org/10.1590/1981-5344/2930.

CURTY, R. Abordagens de reúso e a questão de reusabilidade dos dados científicos. **Liinc em Revista**, Rio de Janeiro, v. 15, n. 2, p. 177-193, nov. 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4777/4315. Acesso em: 9 set. 2021.

DATAONE. **Data life cycle**. Disponível em: https://old.dataone.org/data-life-cycle. Acesso em: 15 ago. 2021.

DATAVERSE. **Harvard Dataverse**. [Cambridge, MA: Harvard College, 2021]. Disponível em: https://dataverse.harvard.edu. Acesso em: 14 ago. 2021.

DATAVERSE + file management. *In*: DATAVERSE PROJECT. **User guide.** [Cambridge: Harvard College], 2016. Disponível em: https://guides.dataverse.org/en/latest/user/dataset-management.html#. Acesso em: 31 ago. 2021.

DIAS, Guilherme Ataíde; ANJOS, Renata Lemos dos; RODRIGUES, Adriana Alves. Os princípios FAIR: viabilizando o reuso de dados científicos. *In*: DIAS, Guilherme Ataíde; OLIVEIRA, Bernardina Maria Juvenal Freire de. (org.). **Dados científicos:** perspectivas e desafios. João Pessoa: Editora UFPB, 2019. cap. 8, p. 177-187. Disponível em: http://www.editora.ufpb.br/sistema/press5/index.php/UFPB/catalog/download/359/508/2949-1?inline=1. Acesso em: 13 ago. 2021.

EMBRAPA. Resolução Consad nº 184, de 4 de abril de 2019. Política de Governança de Dados, Informação e Conhecimento da Embrapa. **Boletim de Comunicações Administrativas,** Brasília, DF, ano 45, n. 16, p. 1-19, 5 abr. 2019. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/212939/1/A-politica-de-governanca.pdf. Acesso em: 01 jul. 2019.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Laboratório Multiusuário de Bioinformática.** Disponível em: https://www.embrapa.br/informatica-agropecuaria/lmb. Acesso em: 14 ago. 2021.

ESTADOS UNIDOS. Department of Agriculture. **NAL agricultural thesaurus and glossary**. Disponível em: https://agclass.nal.usda.gov/. Acesso em: 15 ago. 2021.

EUROPEAN COMMISSION EXPERT GROUP ON FAIR DATA. **Turning FAIR data into reality:** final report and action plan from the European Commission Expert Group on FAIR Data. Brussels,

2018. 76 p. Disponível em: https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf. Acesso em: 11 ago. 2021.

FAO. **AGROVOC**. Disponível em: http://aims.fao.org/agrovoc. Acesso em: 8 jul. 2021.

THE FUTURE OF RESEARCH COMMUNICATIONS AND E-SCHOLARSHIP. **About FORCE 11**. Disponível em: https://www.force11.org/about. Acesso em: 12 ago. 2021a

THE FUTURE OF RESEARCH COMMUNICATIONS AND E-SCHOLARSHIP. **The FAIR data principles.** Disponível em: https://www.force11.org/group/fairgroup/fairprinciples. Acesso em: 12 ago. 2021b.

GERHARDT, Tatiana Engel; RAMOS, Ieda Cristina Alves; RIQUINHO, Deise Lisboa; SANTOS, Daniel Labernarde dos. Estrutura do projeto de pesquisa. In: GERHARDT, Tatiana Engel; SILVEIRA, Denise Tolfo. (org.). **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009. p. 65–88. (Série educação a distância).

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 6. ed. São Paulo: Atlas, 2017.
GO FAIR. **FAIR principles**. [2016]. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 27 ago. 2020.

GORDANA, Rudić; DRAGAN, Ivanović. Cataloguing dataset in Library Information Systems using the MARC 21 format. *In:* INTERNATIONAL CONFERENCE ON INFORMATION SOCIETY AND TECHNOLOGY, 7., 2017, Kopaonik. **Proceedings [...]**. Belgrade: Society for Information Systems and Computer Networks Belgrade, 2017. p. 395-399. Disponível em: http://www.eventiotic.com/eventiotic/files/Papers/URL/379ec28c-39de-44c7-90ba-33f9cbace951.pdf. Acesso em: 28 ago. 2020.

HENNING, Patrícia Corrêa; RIBEIRO, Claudio Jose Silva; SALES, Luana Faria; MOREIRA, Luiz Rebelo; SANTOS, Luiz Olavo Bonino da Silva. Desmistificando os princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados FAIR. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, João Pessoa, v. 14, n. 3, p. 175-192, 2019. Disponível em: https://periodicos.ufpb.br/index.php/pbcib/article/view/46969/27455.

MEY, Eliane Sayão Alves. A. **Introdução à catalogação**. Brasília, DF: Briquet de Lemos, 1995. 123 p.

MEY, Eliane Sayão Alves; SILVEIRA, Naira Christofoletti. **Catalogação no plural.** Brasília, DF: Briquet de Lemos, 2009. 217 p.

MOREIRA, Fábio Mosso; SANT'ANA, Ricardo César; SANTOS, Plácida Leopoldina Ventura Amorim da Costa; ZAFALON, Zaira Regina. Metadados para descrição de datasets e recursos informacionais do "Portal Brasileiro de Dados Abertos". **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 22, n. 3, p. 158-185, jul./set. 2017. DOI: https://doi.org/10.1590/1981-5344/2947.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (Estados Unidos). **Taxonomy**. Bethesda. Disponível em: https://www.ncbi.nlm.nih.gov/taxonomy. Acesso em: 10 jul. 2021.

OLIVEIRA, Adriana Carla Silva de; SILVA, Edilene Maria da. Ciência aberta: dimensões para um novo fazer científico. **Informação & Informação**, Londrina, v. 21, n. 2, p. 5-39, maio/ago. 2016. Disponível em:

https://www.uel.br/revistas/uel/index.php/informacao/article/view/27666/20113. Acesso em: 6 ago. 2021. DOI: http://dx.doi.org/10.5433/1981-8920.2016v21n2p5.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **OECD principles and guidelines for access to research data from public funding.** Paris, 2007. 24 p. Disponível em: https://www.oecd.org/sti/inno/38500813.pdf. Acesso em: 10 ago. 2021.

ROCHA, Lucas de Lima; SALES, Luana Faria; SAYÃO, Luís Fernando. Descrever para preservar: metadados como ferramenta para gestão de dados de pesquisa. **ISKO Brasil,** *[s. l.]*, v. 5, p. 194-201, 2017. Disponível em: https://www.brapci.inf.br/index.php/res/v/121924. Acesso em: 12 ago. 2021.

ROCHA, Rafael Port da; GABRIEL JUNIOR, Rene Faustino; VANZ, Samile Andréa de Souza; BORGES, Eduardo Nunes; AZAMBUJA, Luís Alberto Barbosa; CAREGNATO, Sônia Elisa; PAVÃO, Caterina Groposo; PASSOS, Paula Caroline Schifino Jardim; FELICISSIMO, Carolina Howard. Análise dos sistemas DSpace e Dataverse para repositórios de dados de pesquisa com acesso aberto. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 17, p. 1-25, 2021. Disponível em: http://hdl.handle.net/20.500.11959/brapci/160963. Acesso em: 1 set. 2021.

SALES, Luana Faria; SAYÃO Luís Fernando; MARANHÃO, Ana Maria Neves; DRUMOND, Geisa Meirelles; SILVA, Maria Helena Ferreira Xavier da. Competências dos bibliotecários na gestão dos dados de pesquisa. **Ciência da Informação**, Brasília, DF, v. 48, n. 3, p. 303-313, set./dez. 2019. Disponível em: http://revista.ibict.br/ciinf/article/view/4973/4458. Acesso em: 7 ago. 2021.

SAYÃO, Luís Fernando; SALES, Luana Faria. **Guia de gestão de dados de pesquisa:** [minicurso]. [Rio de Janeiro: CNEN, 2016]. 196 slides.

SEMELER, Alexandre Ribas; PINTO, Adilson Luiz. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia de dados. **Ciência da Informação**, Brasília, DF, v. 48, n. 1, p. 113-129, jan./abr. 2019. Disponível em: http://revista.ibict.br/ciinf/article/view/4461/4102. Acesso em: 8 ago. 2021.

SILVA, Rhanna Henriques Guimarães da; SILVEIRA, Naira Christofoletti. Considerações sobre catalogação de cervejas artesanais. **Biblionline,** João Pessoa, v. 13, n. 2, p. 102-115, abr./jun. 2017. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/biblio/article/view/35685/18830. Acesso em: 15 ago. 2021.

SOUZA, Márcia Izabel Fugisawa; VISOLI, Marcos Cezar; TORRES, Tércia Zavaglia. **Catalogação de dataset no Repositório de Dados da Embrapa:** a experiência do projeto-piloto de implantação de gestão de dados de pesquisa no Laboratório Multiusuário de Bioinformática. Campinas: Embrapa Informática Agropecuária, 2020. 117p. (Embrapa Informática Agropecuária. Documentos, 172). Disponível em: https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1127941/1/Doc172-2020.pdf. Acesso em: 28 dez. 2020.

VEIGA, Viviane. Gestão de dados de pesquisa FAIR: dando um Jump em seus dados. *In*: ENCONTRO DA REDE SUDESTE DE REPOSITÓRIOS INSTITUCIONAIS, 1., 2019, Rio de Janeiro. **Anais [...].** Rio de Janeiro: Fiocruz/Icict/UFRJ, 2019. 59 p. Disponível em: https://www.arca.fiocruz.br/bitstream/icict/33343/2/ve_Veiga_Viviane_ICICT_I_Encontro_Sudeste_RIAA_2019.pdf. Acesso em: 15 ago. 2021.

VOCABULÁRIO controlado. *In:* CUNHA, Murilo Bastos da; CAVALCANTI, Cordélia Robalinho de Oliveira. **Dicionário de biblioteconomia e arquivologia.** Brasília, DF: Briquet de Lemos, 2008. p. 378.

WILKINSON, Mark D.; DUMONTIER, Michel; AALBERSBERG, IJsbrand Jan; APPLETON, Gabrielle; AXTON, Myles; BAAK, Arie; BLOMBERG, Niklas; BOITEN, Jan-Willem; SANTOS, Luiz Bonino da Silva; BOURNE, Philip E.; BOUWMAN, Jildau; BROOKES, Antony J.; CLARK, Tim; CROSAS, Mercè; DILLO, Ingrid; DUMON, Olivier; EDMUNDS, Scott; EVELO, Chris T.; FINKERS, Richard; GONZALEZ-BELTRAN, Alejandra; GRAY, Alasdair J. G.; GROTH, Paul; GOBLE, Carole; GRETHE, Jeffrey S.; HERINGA, Jaap; HOEN, Peter A. C 't; HOOFT, Rob; KUHN, Tobias; KOK, Ruben; KOK, Joost; LUSHERM, Scott J.; MARTONE, Maryann E.; MONS, Albert; PACKER, Abel L.; PERSSON, Bengt; ROCCA-SERRA, Philippe; ROOS, Marco; SCHAIK, Rene van; SANSONE, Susanna-Assunta; SCHULTES, Erik; SENGSTAG, Thierry; SLATER, Ted; STRAWN, George; SWERTZ, Morris; THOMPSON, Marke; LEI, Johan van der; MULLIGEN, Erik van; VELTEROP, Jan; WAAGMEESTER, Andra; WITTENBURG, Peter; WOLSTENCROFT, Katherine; ZHAO, Jun; MONS, Barend. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data,** London, v. 3, p. 1-9, 2016. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 3 set. 2021.

Recebido em: 29 de setembro de 2021
Aprovado em: 30 de abril de 2022
Publicado em: 17 de junho de 2022