Análise dos sistemas DSpace e Dataverse para repositórios de dados de pesquisa com acesso aberto

Analysis of DSpace and Dataverse systems for open access research data repositories

Rafael Port da Rocha
Doutor em Computação pela Universidade Federal do Rio Grande do Sul (UFRGS). Professor associado da Universidade Federal do Rio Grande do Sul (UFRGS).
rafael.rocha@ufrgs.br

Samile Andréa de Souza Vanz
Doutora em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul (UFRGS). Professora associada da Universidade Federal do Rio Grande do Sul (UFRGS).
samilevanz@terra.com.br

Luís Alberto Barbosa Azambuja
Mestre em Engenharia de Computação pela Universidade Federal do Rio Grande (FURG). Técnico em Tecnologia da Informação na Universidade Federal do Rio Grande (FURG).
l.azambuja@gmail.com

Caterina Groposo Pavão
Doutora em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul (UFRGS). Professora na Universidade Federal do Rio Grande do Sul (UFRGS).
caterina@cpd.ufrgs.br

Carolina Howard Felicissimo
Doutora em Informática pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-RJ). Coordenadora de Pesquisa e Desenvolvimento da Rede Nacional de Ensino e Pesquisa.
carolina.felicissimo@rnp.br

Rene Faustino Gabriel Junior
Doutor em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). Professor adjunto da Universidade Federal do Rio Grande do Sul (UFRGS).
renefgj@gmail.com

Eduardo Nunes Borges
Doutor em Computação pela Universidade Federal do Rio Grande do Sul (UFRGS). Professor adjunto da Universidade Federal do Rio Grande (FURG).
eduardoborges@furg.br

Sônia Elisa Caregnato
Doutora em Information Studies pela University of Sheffield. Professora titular da Universidade Federal do Rio Grande do Sul (UFRGS).
sonia.caregnato@ufrgs.br

Paula Caroline Schifino Jardim Passos
Doutora em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul (UFRGS). Professora substituta da Universidade Federal do Rio Grande do Sul (UFRGS).
paulacarolinejardim@gmail.com

RESUMO

Este artigo apresenta a análise comparativa dos softwares DSpace e Dataverse para compartilhamento de dados abertos de pesquisa, realizada no âmbito do projeto RDP Brasil. A análise das soluções tecnológicas enfatiza as funcionalidades que cada software oferece para dar apoio à construção de um repositório de dados de pesquisa que venha a prover o compartilhamento de dados segundo os princípios FAIR e que atenda a critérios estabelecidos para repositórios digitais confiáveis. Os sistemas são analisados a partir de critérios do modelo OAIS classificados em sete requisitos: representação do ambiente do repositório; representação dos conjuntos de dados; descrição e documentação dos conjuntos de dados; produção dos conjuntos de dados; armazenamento a longo prazo e planejamento da preservação; acesso e uso dos conjuntos de dados; e uso, desenvolvimento e manutenção do software. As soluções tecnológicas DSpace e Dataverse foram investigadas em profundidade por serem as mais comumente adotadas por repositórios digitais confiáveis. Conclui-se que o Dataverse possui recursos para configuração de vários tipos de ambientes de repositório, incluindo hierarquias organizacionais e políticas de gestão distintas para

unidades ou grupos, com esquemas de metadados e licenças. O DSpace também permite tais configurações, porém, por ter sido desenvolvido para repositórios institucionais de documentos e por estar estruturado a partir do conceito de coleção de itens, adaptações são necessárias.

Palavras–chave: Acesso aberto a dados de pesquisa; Software; Repositório de dados de pesquisa; DSpace; Dataverse.

ABSTRACT

This paper presents a comparative analysis of DSpace and Dataverse software to share open research data carried out by RDP Brasil Project. An analysis of technological solutions emphasizes the functionalities that each software offers to support the creation of a research data repository that provides data sharing according to FAIR principles and that comply with trustworthy digital repositories. The systems are analyzed using OAIS model criteria classified into seven requirements: representation of repository environment; datasets representation; description and documentation of datasets; datasets creation; long term storage and preservation planning; access and use of datasets; and use, development and maintenance of the software. DSpace and Dataverse were investigated in-depth because they are the most common software adopted by digital repositories. We conclude that Dataverse has features for configuring various types of repository environments, including organizational hierarchies and distinct management policies for units or groups, with metadata and statistics schemes. DSpace also allows such configurations, with adaptations, because it was developed for institutional document repositories, and it is structured based on the concept of item collection.

Keywords: Research data open access; Software; Research data repository; DSpace; Dataverse.

1 INTRODUÇÃO

A ciência aberta tem entre um dos seus pressupostos o compartilhamento de dados de pesquisa em Repositórios de Dados (ALBAGLI; CLINIO*;* RAYCHTOCK, 2014). Esses repositórios podem ser disciplinares (quando armazenam dados que seguem padrões e recomendações de disciplinas específicas), institucionais, nacionais ou multidisciplinares (caraterizados como repositórios de cauda longa por armazenarem dados com características das mais diversas áreas de pesquisa) (SAYÃO; SALES, 2015).

Considerando o ciclo de vida do dado, repositórios podem armazenar dados de pesquisas concluídas ou envolver dados de pesquisas em andamento, quando os conjuntos de dados podem ser atualizados, resultando em versões diferentes versões de um mesmo conjunto de dados, existem também os repositórios de dados associados à publicação de artigos (SAYÃO; SALES, 2012).

Os repositórios também podem ter diversas políticas de submissão e estratégias para organização dos materiais submetidos. Um repositório pode possuir como política armazenar somente conjuntos de dados resultantes de estudos, em uma estrutura plana, não hierárquica. Por outro lado, um repositório pode ter como estratégia permitir que grupos de pesquisa armazenem os conjuntos de dados de seus vários estudos; ou que

organizações disponibilizem seus grupos e os estudos produzidos por esses grupos (SAYÃO; SALES, 2018).

A implantação de um repositório de dados é orientada por uma série de modelos de referência e por princípios, como o Modelo de Referência para Repositórios - *Open Archival Information System* (OAIS) (THE CONSULTATIVE COMMITTEE..., 2012) e os princípios de compartilhamento FAIR (WILKINSON *et al.*, 2016; HENNING et al., 2019) e de citação (DATA CITATION..., 2014), que indicam que dados devem ser localizáveis, acessíveis, interoperáveis e reusáveis. Além disso, para o desenvolvimento de um repositório, é importante considerar critérios para repositórios digitais confiáveis e preservação digital, visando a obtenção de certificações, como ISO 16363 (INTERNATIONAL STANDARD ORGANIZATION, 2012), Data Seal Approval (LEEUW, 2019), Core Trust Seal[1] e Nestor[2].

Este estudo tem como objetivo investigar soluções tecnológicas para repositórios de dados disponíveis no mercado com software de código aberto. O trabalho adota como estratégia analisar de que forma estes softwares proporcionam instrumentos que permitam o desenvolvimento de repositórios digitais confiáveis, de acordo com princípios FAIR (WILKINSON *et al.*, 2016; HENNING et al., 2019), com padrões estabelecidos para citação, proveniência e certificação e com o modelo de Referência OAIS.

O artigo resulta do projeto de pesquisa RDP Brasil - Rede de Dados de Pesquisa Brasileira (GABRIEL JUNIOR et al., 2019), desenvolvido pelo Centro de Documentação e Acervo Digital de Pesquisa (CEDAP), e o Centro de Processamento de Dados (CPD), ambos da Universidade Federal do Rio Grande do Sul (UFRGS), e o Grupo de Pesquisa em Gerenciamento de Informações do Centro de Ciências Computacionais (C3), da Universidade Federal do Rio Grande (FURG).

As próximas seções apresentam os procedimentos metodológicos adotados, seguidos pelos resultados e considerações finais.

2 PROCEDIMENTOS METODOLÓGICOS

Para investigar as soluções tecnológicas foram considerados repositórios que fossem capazes de permitir a representação e a configuração de políticas de submissão,

[1] Core Trust Seal - https://www.coretrustseal.org/
[2] Nestor - http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html

acesso e gestão para conjuntos de dados, estudos e grupos. Também considerando que o repositório deve ser capaz de representar tanto dados produzidos durante a pesquisa, como dados a serem preservados após o seu término, além de permitir o gerenciamento de versões de dados, decorrentes de transformações que ocorrem no ciclo de vida do dado, tanto dados coletados e processados no decorrer da pesquisa, como dados para serem reutilizados.

A escolha dos softwares DSpace e do Dataverse para a análise, foi baseada nas informações disponíveis no diretório de repositórios de dados Re3Data[3], tendo como critério os repositórios que obtiveram a certificação de repositório confiável, e que estivessem disponíveis em acesso aberto. Também foram identificados 14 repositórios em Fedora, sete em Nesstar, quatro em MySQL e 16 em outras plataformas, e 33 que não informaram o software utilizado.

As principais soluções tecnológicas usadas para repositórios de dados de pesquisa cadastradas no diretório de repositórios de dados Re3Data e que atendem a característica de serem distribuições completas de software para repositórios são os softwares DSpace, Dataverse, CKAN, Fedora e EPrints. Este artigo restringe-se ao DSpace e ao Dataverse, por serem as soluções mais utilizadas e adotadas por repositórios que obtiveram certificação de repositório confiável. Em novembro de 2018 o Re3Data registrava 62 repositórios em DSpace e 69 Dataverse, em maio de 2021 estavam cadastrados 98 repositórios em DSpace e 105 em Dataverse. A investigação dos softwares foi feita a partir da sua documentação, além de relatos encontrados na literatura.

Para analisar as soluções tecnológicas foram elaborados 56 critérios com base no modelo OAIS (THE CONSULTATIVE COMMITTEE..., 2012) e para identificar a interface entre produtor e a base de dados (THE CONSULTATIVE COMMITTEE..., 2012). Esses critérios, que foram apresentados inicialmente em Rocha e outros autores (2018), estão classificados em sete categorias, e cada categoria foi relacionada com critérios para certificação de repositório confiável de Core Trust Seal e com os princípios FAIR. O Quadro 1 resume os critérios de cada categoria.

[3] Diretório de Repositórios de Dados Re3data - https://www.re3data.org/browse

Quadro 1: Critérios para análise das soluções tecnológicas

Categoria	Critérios
Desenvolvimento, Manutenção e Uso do Software	Customização da Interface, Usabilidade, Tecnologia e Plataforma Distribuição e Versionamento, Estratégia de Desenvolvimento do Software Licença de Uso, Desempenho e Escalabilidade, Presença – Usuários, Uso no Brasil.
Representação do Ambiente do Repositório	Prover meios para representar o ambiente do repositório (como comunidades, subcomunidades, coleções, grupos), políticas de funcionamento (autorizações, grupos, papéis, grupos, fluxos), gestão descentralizada, integração com Current Research Information Systems (CRIS) e web semântica.
Representação dos Conjuntos de Dados	Prover meios para representar um conjunto de dados (pacotes), considerando estruturas, versões, formatos de arquivos e empacotamento
Descrição e Documentação dos Conjuntos de Dados	Prover meios para produzir, representar e gerenciar metadados ricos, precisos, indexáveis e compreensíveis por máquina. Criação de esquemas, registro de informação descritiva, administrativa e de representação e sua representação em metadados. Documentar os conjuntos de dados.
Produção dos Conjuntos de Dados	Prover meios para submissão dos dados ao repositório, observando pacotes de submissão, fluxo de submissão, licença, funções que verificam a autenticidade e autoridade do produtor, a integridade do material submetido (verificação) e a conformidade (validação) com as especificações planejadas.
Armazenamento a Longo Prazo e Planejamento da Preservação	Prover meios para armazenamento seguro da informação em pacotes de armazenamento e realização de ações de preservação digital
Acesso e Uso dos Conjuntos de Dados	Prover meios para descoberta dos dados; restringir o acesso a dados a pessoas ou grupos autorizados; entregar dados ao consumidor em formatos usados por estes; prover acesso aos dados, seus metadados e outras informações.

Fonte: Dados da pesquisa apresentados em Rocha *et al.* (2018)

A seção a seguir apresenta as avaliações dos softwares DSpace e Dataverse segundo os critérios para análise de soluções tecnológicas definidos no Quadro1.

3 CARACTERÍSTICAS DOS SOFTWARES DSPACE E DATAVERSE

O software DSpace foi lançado como um esforço conjunto entre os desenvolvedores do MIT Libraries (do Massachusetts Institute of Technology) e do HP Labs (Hewlett-Packard). Com o crescimento da comunidade de usuários, HP e MIT formaram a DSpace Foundation, uma organização sem fins lucrativos que ofereceu suporte aos usuários do software por dois anos a partir de 2007. Desde 2009, a comunidade de usuários é organizada pela DuraSpace, organização criada a partir da

colaboração entre DSpace Foundation e Fedora Commons. Atualmente, DSpace é disponibilizado para plataformas baseadas nos sistemas operacionais UNIX-like (Linux, HP/UX, Mac OSX) ou Microsoft Windows (DURASPACE, 2020).

O DSpace adota o esquema de distribuição de versionamento do software onde as versões principais (*major releases*) podem incluir novas funcionalidades, melhorias de sistema, mudanças arquiteturais e correção de falhas, e versões menores (*minor releases*) são geradas a partir de correção de falhas (*bug fixes*) em versões principais. A política de suporte fornece atualizações de segurança para as três mais recentes versões principais, entretanto apenas a mais recente recebe atualização para correção de falhas.

Distribuído sob licença Berkeley Software Distribution (BSD), o DSpace pode ser redistribuído com ou sem modificação, mas usa muitas bibliotecas de terceiros regidas por diferentes tipos de licenças que devem ser observadas (DURASPACE, 2020). Características de desempenho e escalabilidade do DSpace são dependentes da configuração do servidor Java web, do sistema gerenciador de banco de dados relacional e do sistema de recuperação de informações.

Existem 136 instalações de DSpace no Brasil registradas na comunidade Duraspace (DURASPACE, 2020), principalmente em instituições acadêmicas e governamentais. Entretanto, a grande maioria destas instalações não são dedicadas a repositórios de dados, mas a documentos. Na comunidade Re3data estão registradas apenas três instalações de repositórios de dados brasileiras.

O software Dataverse, está sendo desenvolvido pelo Institute for Quantitative Social Science (IQSS) de Harvard, junto com muitos colaboradores de todo o mundo. O Projeto Dataverse foi construído com base na experiência prévia no projeto anterior Virtual Data Center (VDC) que durou entre 1999 e 2006[4]. De acordo com Crosas (2011), o principal objetivo da Rede Dataverse é resolver os problemas de compartilhamento de dados por meio da criação de tecnologias que permitam às instituições reduzir a carga de trabalho de pesquisadores e editores de dados e incentive-os a compartilhar seus dados.

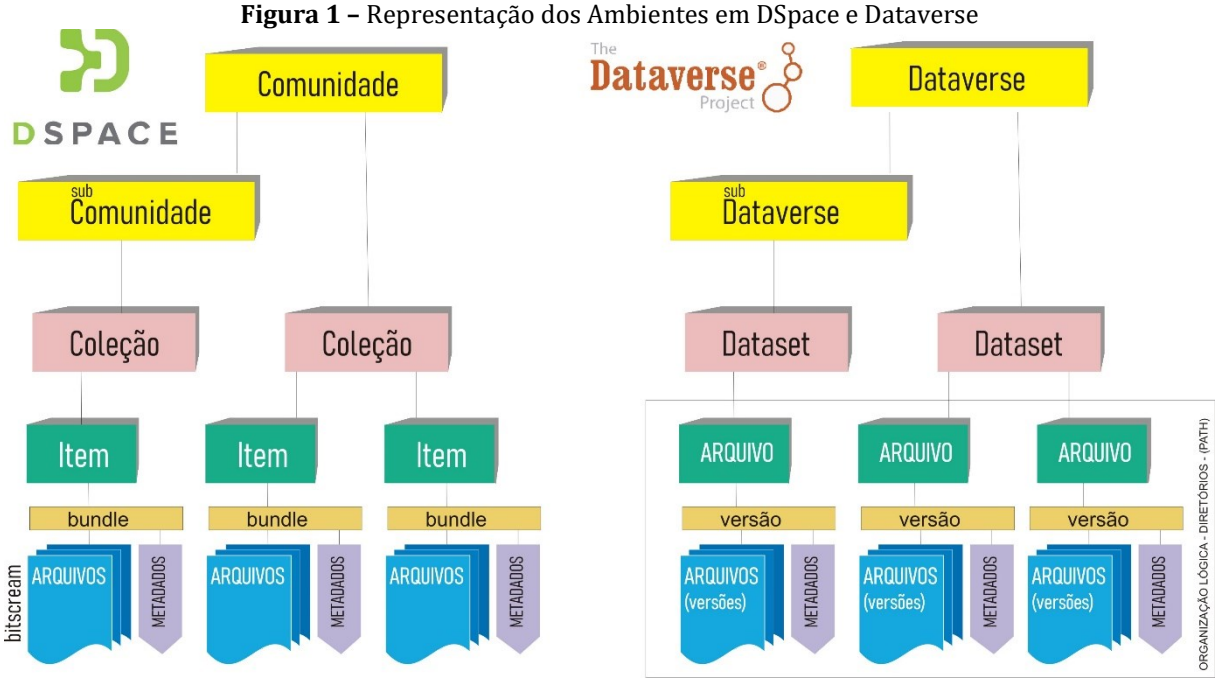
O esquema de distribuição e versionamento do Dataverse inclui versões principais (*major releases*) com novas funcionalidades, mudanças arquiteturais ou de sistema. Sua distribuição é realizada sob licença Apache 2.0[5], permitindo ao usuário a liberdade de usar

[4] Dataverse - https://github.com/IQSS/dataverse/releases
[5] Licença Dataverse - https://github.com/IQSS/dataverse/blob/develop/LICENSE.md

o software para qualquer finalidade, distribuí-lo, modificá-lo e distribuir versões modificadas, sob os termos da licença, sem preocupação com *royalties*.

Em relação à representação do ambiente dos repositórios, conforme critérios estabelecidos no quadro 1, apresentamos, na figura 1, os elementos (nomenclatura) do DSpace e do Dataverse, conforme sua estrutura do ambiente.

Figura 1 – Representação dos Ambientes em DSpace e Dataverse



Fonte: Elaborado pelos autores (2021).

O DSpace disponibiliza recursos que permitem variadas possibilidades para ambientes de repositório. Comunidades são os recursos do DSpace para representar entidades organizacionais, como unidades, subunidades e/ou grupos. As comunidades podem conter subcomunidades e coleções. Uma coleção compreende um conjunto de itens com características similares, descrito por um mesmo esquema de metadados, que está sujeito ao mesmo fluxo e às mesmas políticas de submissão, de gestão e de acesso

Apesar do software não ter sido desenvolvido para dados de pesquisa, as comunidades podem ser usadas para representar instituições produtoras ou custodiantes de conjuntos de dados, e/ou grupos de pesquisa. Por exemplo, o DataShare, repositório de dados da Universidade de Edimburgo, usa comunidades e subcomunidades para representar hierarquicamente as unidades da Universidade[6].

[6] Comunidades de Datashare - https://datashare.is.ed.ac.uk/community-list

Com relação a políticas de funcionamento e gestão descentralizada, o DSpace dispõe de recursos para implementar comunidades e coleções com políticas próprias e distintas, por meio da organização de usuários em grupos. O software permite a criação de grupos com autorizações para acessar comunidades, coleções e itens, assim como para submeter itens e para criar e gerenciar comunidades, coleções, fluxos de submissões e itens.

O Dataverse, da mesma forma, possui recursos que possibilitam variadas configurações de ambientes de repositório de dados. A entidade dataverse é a estrutura que o software disponibiliza para representar organizações, grupos ou unidades. O software também é capaz de representar estruturas organizacionais hierárquicas, à medida que dataverses podem conter outros dataverses. Cada entidade dataverse contém *datasets*, que são as entidades que representam conjuntos de dados. O software dispõe de recursos para implementar entidades dataverses com políticas de funcionamento próprias e distintas. Uma entidade dataverse pode ser definida como se fosse um repositório independente, com todas autorizações para gerenciamento e operação (como definir papéis, permissões e grupos; criar e gerenciar entidades dataverses, *datasets*, arquivos etc.). Cada entidade dataverse pode ter marca personalizada (interface, logotipo) e interface configurada para descoberta de dados.

Para a definição de políticas, o software trabalha os conceitos permissão, papel e grupo. Permissão representa a autorização para realização de uma determinada funcionalidade do ambiente, como publicar um conjunto de dados. Papel corresponde a um conjunto de permissões, que caracteriza um perfil exercido por usuários na gestão ou operação do repositório. O ambiente já disponibiliza vários papéis (Administrador, Publicador de Dataverse, Curador de Dataverse, Editor de *Dataset*, etc.), mas novos papéis podem ser criados. Grupo representa um conjunto de usuários habilitados a atuar com determinados papéis.

No DSpace, um conjunto de dados pode ser representado como um item de uma coleção, que é uma estrutura complexa, composta por *bundles* (espécies de pastas) e *bitstreams*, que são arquivos binários. O DSpace permite a representação de arquivos nos mais variados formatos, como texto (txt, pdf, csv), imagem (jpg e tiff), vídeo (mpeg) e som (wave). Os formatos planejados como aceitos são cadastrados no módulo chamado Registro de Formato.

Um módulo do DSpace foi criado para prover serviço de versionamento de item, sendo incorporado na distribuição de DSpace (DSPACE 6.x DOCUMENTATION, 2018). Para cada nova versão de um item, um item separado será criado, que replica os registros de metadados, *blundle* e *bitstreams*. Os registros dos *bitstreams* irão apontar para os mesmos arquivos do disco. Por exemplo, no Repositório Dryad[7], a atualização de algum arquivo de um item leva a uma nova versão, e cada arquivo dessa versão possui um novo identificador.

Um fator importante para a preservação digital é representar, para fins de armazenamento em longo prazo, os itens em pacotes e não somente em estruturas próprias do software do repositório. O DSpace representa itens de forma dependente do software. Arquivos (*bitstreams*) são armazenados no sistema de arquivos, e a estrutura do item (seus metadados e *bundles*) são representados em banco de dados relacional do DSpace. Isso torna a obtenção do item dependente do software, trazendo dificuldades para desenvolver estratégias de preservação digital. Entretanto, para superar essa limitação o DSpace permite a importação/exportação de pacotes de armazenamento de informação em estruturas que seguem padrões e recomendações de preservação digital. Pacotes são representados por estruturas de pastas, com o uso do padrão METS para descrever metadados estruturais (DSPACE 6.x DOCUMENTATION, 2018).

No Dataverse, grupos de pesquisa ou entidades (unidades, órgãos) são representados por de entidades dataverse. Cada estudo realizado por um grupo ou entidade compreende um conjunto de dados (*dataset)*. Um *dataset* é composto por metadados, pelos termos de uso (como licenças) e por arquivos. Metadados descrevem o *dataset* e os arquivos. As estruturas de um *dataset* são direcionadas para representar dados de pesquisa, sendo compatíveis com padrão de metadados DDI Lite[8] e DDI 2.5 Codebook[9]. Esses esquemas de metadados visam descrever dados de pesquisa no contexto de um estudo. No caso de conjuntos de dados tabulares, ocorre também a descrição das estruturas das variáveis representadas nessas tabelas.

No Dataverse, documentos e conjuntos de dados podem ser representados por diferentes formatos, a fim de atender demandas de variados consumidores. O ambiente utiliza a ferramenta JHOVE para identificar, validar e descrever o formato de um arquivo

[7] Repositório Dryad - https://datadryad.org/
[8] DDI Lite – Elementos recomendados de DDI Codebook http://www.ddialliance.org/sites/default/files/ddi-lite.html
[9] DDI Codebook 2.5 - http://www.ddialliance.org/Specification/DDI-Codebook/2.5/

submetido. Entretanto, esse ambiente não disponibiliza funcionalidades para permitir o estabelecimento de políticas e de controle de formatos aceitos. O ambiente possui recursos especiais para tratar arquivos em formatos tabulares e geoespaciais. Com relação a arquivos submetidos com dados tabulares (CSV, Excel, SPSS, R etc.), o Dataverse gera representações em formato uniforme e aberto (.tab), para permitir que os arquivos possam ser usados uniformemente por ferramentas externas (como R) e por ferramentas já integradas com o ambiente, como TwoRavens[10] (dados tabulares) e WordMap[11] (dados geoespaciais). O Dataverse confere a integridade e armazena dados geoespaciais em formatos vetoriais (*shapefiles*), a fim de permitir sua integração com a ferramenta WordMap e quaisquer Sistemas de Informação Geográfica (GIS).

O ambiente permite o versionamento de *datasets*. Cada alteração nos metadados ou nos arquivos (adição, remoção) leva à criação de uma nova versão, com identificação global persistente e citação para essas versões. Edições mínimas em metadados levam a novas versões intermediárias sem que a citação seja alterada.

Em Dataverse, um *dataset* tem as informações de sua estrutura, seus termos e seus metadados armazenados no banco de dados relacional Postgres de Dataverse. Já os arquivos deste dataset são armazenados em sistema de arquivos. Embora represente *datasets* em estruturas dependentes do software, o Dataverse permite a exportação dos metadados de um *dataset* no formato DDI Codebook, que resulta em um arquivo Extensible Markup Language (XML) que descreve todo o pacote, incluindo metadados estruturais (estruturas físicas e lógicas dos documentos, como variáveis em documentos tabulares). O Dataverse também permite que metadados e arquivos que compõem *datasets* sejam extraídos do ambiente via interface de programação (API), facilitando a preservação a longo prazo.

Quanto a documentação dos softwares, o DSpace permite a definição de esquemas de metadados, com elementos não hierárquicos (elementos não podem conter outros elementos). O software representa informações de formato e proveniência/preservação digital por meio de metadados, e representa metadados estruturais do pacote e outras informações na base de dados relacional do ambiente. Para descrever itens, permite a criação de novos elementos, e a organização desses elementos em novos esquemas. Dessa

[10] TwoRavens - http://2ra.vn/
[11] Worldmap - http://worldmap.harvard.edu/

forma, o software possibilita a definição de esquemas de metadados que atendam às necessidades da comunidade do repositório.

No DSpace, metadados são representados na forma propriedade-valor. Por isso, o software não permite a representação de elementos com estruturas cujos conteúdos são também desdobrados em novos elementos. Entretanto, ele dispõe de mecanismos de mapeamento de metadados, via regras eXtensible Stylesheet Language for Transformation (XSLT), que permitem a geração de representações de metadados em estruturas hierárquicas XML a partir de representações planas, propriedade-valor.

Com relação a metadados descritivos, o DSpace tem como base o esquema Dublin Core[12]. Dispõe, em sua versão de instalação, dois esquemas: Dublin Core Qualificado (DSPACE 6.x DOCUMENTATION, 2018) e Dublin Core Terms[13], que podem ser estendidos, como realizado no repositório DataShare[14].

Com relação a metadados administrativos, o DSpace possui um fluxo para submissão de item e possibilita também a submissão por máquina. As ações realizadas na submissão são armazenadas como metadados de proveniência (elemento dc.description.provenance). São registrados o arquivo, seu *checksum*, data e usuário que executou a ação.

Com relação aos metadados administrativos técnicos, na submissão de um arquivo que compõe um item, o DSpace identifica o formato do arquivo submetido (pela terminação indicada no nome do arquivo) e registra essa informação por meio do metadado de Dublin Core dc.format. A ferramenta também registra o texto descritivo informado e *checksum,* calculado.

Com relação a metadados estruturais, a descrição do pacote de informação, isto é, do pacote que contém o item (*bundles*, *bitstreams*, metadados) é representada no banco de dados relacional que dá suporte ao funcionamento do repositório. Conforme observado anteriormente, o DSpace não armazena o item em um único local (sistema de arquivos), de forma independente do software, mas permite a exportação de pacotes, e essa exportação segue padrões de estruturas e metadados recomendados para preservação digital[Erro! Indicador não definido.]. Em um pacote exportado pelo DSpace, os

[12] Esquema de Metadados Dublin Core - http://dublincore.org/documents/dces/
[13] Esquema Dublin Core Terms - http://dublincore.org/documents/dcmi-terms/
[14] Metadados do Repositório DataShare - https://www.wiki.ed.ac.uk/display/datashare/metadata

metadados estruturais são representados no padrão METS[15], a partir das informações de estrutura que estão armazenadas na base de dados relacional. Nesse pacote de exportação, outros metadados administrativos e técnicos também são codificados a partir de informações que estão armazenadas na base de dados, incluindo informações de coleções e grupos de usuários. Esses metadados são representados nos padrões METS e PREMIS[16].

O DSpace não dispõe de interface para criação e gestão de vocabulários controlados. Entretanto, dentre os recursos disponíveis para a definição de formulários de entradas de dados, o sistema possibilita que dados sejam controlados por vocabulários fornecidos em arquivos XML (DSPACE 6.x DOCUMENTATION, 2018). O software dispõe de um recurso que usa o indexador Solr para controlar o preenchimento de metadados de autoridade. As informações desse índice provêm dos valores dos metadados que contêm autoridades dos itens armazenados do DSpace. Esse recurso também pode ser associado com um serviço que busca nomes de autoridades do serviço global de controle de autoridades, o Open Researcher and Contributor ID (ORCID) (DSPACE 6.x DOCUMENTATION, 2018). Nesse serviço, as autoridades são armazenadas e gerenciadas em um ambiente externo ao DSpace, isto é, em um índice de autoridades gerenciado pela ferramenta de indexação Solr.

O DSpace oferece recursos para especificar o mapeamento entre esquemas de metadados (*crosswalks*), para permitir a colheita de metadados em diversos esquemas via protocolo OAI-PMH. O mapeamento é realizado por meio de regras XSLT (DSPACE 6.x DOCUMENTATION, 2018). Em sua instalação padrão são disponibilizadas regras XSLT para permitir a colheita de metadados em esquemas como RDF, METS, MODS[17] e MARC[18].

Não estão disponíveis recursos específicos para gerenciar a documentação que dá apoio ao uso dos dados (como *codebooks*/livros de códigos). Entretanto, é possível configurar uma estrutura de item (via *bundle* e *bitstreams*) que representa arquivos de documentação. Também é possível registrar essas informações em metadados específicos, a serem criados.

[15] METS - Esquema para Metadados Estruturais - Metadata Encoding and Transmission Standard - http://www.loc.gov/standards/mets/
[16] PREMIS - Esquema para Medadados de Preservação Digital Eshttps://www.loc.gov/standards/premis/
[17] Metadata Object Description Schema (MODS) - http://www.loc.gov/standards/mods/
[18] MARC - https://www.loc.gov/marc/

Já o Dataverse é um software de repositório que permite a definição de esquemas de metadados quaisquer. Entretanto, disponibiliza esquemas de metadados pré-definidos, apropriados para as áreas: Ciências Sociais e Humanidades, Astronomia e Astrofísica e Ciências da Vida. A especificação de cada um desses esquemas foi feita visando a compatibilidade com esquemas padrões. Dessa forma, o software disponibiliza mapeamentos (*croswalks*) que permitem que os metadados sejam exportados em representações nesses esquemas padrões. O quadro 2 apresenta os esquemas mapeados.

Quadro 2 – Esquemas de metadados de Dataverse e esquemas padrões compatíveis

Esquema de Metadados de Dataverse	Compatibilidade com esquemas padrões
Metadados de citação	DDI Lite, DDI 2.5 Codebook DataCite 3.1 Dublin Core's DCMI Metadata Terms
Metadados Geoespacial	DDI Lite, DDI 2.5 Codebook DataCite Dublin Core
Metadados de Ciências Sociais e Humanidades	DDI Lite, DDI 2.5 Codebook Dublin Core
Metadados de Astronomia e Astrofísica	International Virtual Observatory Alliance.VOResource
Metadados de Ciências da Vida	ISA-Tab Specification

Fonte: Dataverse Users Guide [19]

O esquema de citação é de uso obrigatório e possui elementos de proveniência, que permitem descrever autor, produtor, colaborador, depositante, fontes, por exemplo. É compatível e exportável para os padrões DDI Lite[20], DDI 2.5 Codebook[9] e Dublin Core. O Dataverse, ao permitir o mapeamento de seus metadados para Dublin Core e para DataCite, dá suporte à descoberta de informações. Ao mapear seus metadados para DDI Codebook, o Dataverse dá suporte à representação de um estudo, a suas variáveis e a arquivos, conforme Ciências Sociais e Humanidades.

O esquema de metadados para Astronomia e Astrofísica permite que os elementos descritos sejam mapeados e exportados para VOResource Schema format[21], esquema desenvolvido para descrever recursos de observações virtuais (*Virtual Observatory*). O esquema de Ciências da Vida de Dataverse é baseado na especificação ISA-Tab[22], que foi definida de acordo com ISA-Framework. Esse esquema envolve a descrição de dados

[19] Dataverse User Guide. Atualização 8/2019. http://guides.dataverse.org/en/latest/user/index.html
[20] DataCite 3.1 - https://schema.datacite.org/meta/kernel-3.1/
[21] VOResource Schema - http://www.ivoa.net/documents/VOResource/20180625/index.html
[22] ISA-Framework - ISA-Tab - https://www.isacommons.org/

experimentais, envolvendo características de amostras, tecnologias e tipos de medição, relações entre amostras e dados.

Metadados que indicam termos de uso dos dados, como licença, são representados no componente "Termos de um *dataset*". O Dataverse também disponibiliza metadados para descrever os arquivos (como descrição, formato de arquivo, *checksum* – MD5, Universal Numeric Fingerprint (UNF) – e variáveis, em caso de estruturas tabulares).

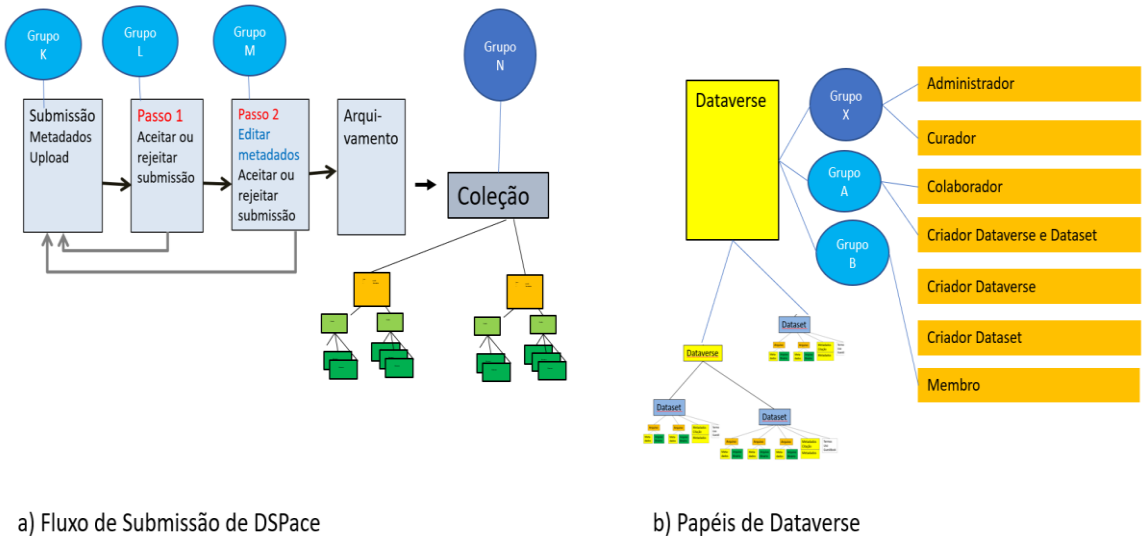
O Dataverse registra ações administrativas relativas à submissão, como quem depositou e a data do depósito, quem publicou e a data do depósito. Ao gerenciar versões, o ambiente permite rastrear todas as mudanças que ocorreram no *dataset* ao longo de seu ciclo de vida, isto é, realiza o registro de ações administrativas referentes à proveniência. O ambiente registra e mostra em interface de usuário as versões de um *dataset*, seus publicadores, data de publicação e mudanças ocorridas. Para arquivos depositados em *datasets*, o Dataverse possibilita a inclusão de informações específicas de proveniência, por meio da descrição textual da proveniência e/ou da inclusão (importação) de um arquivo com metadados de proveniência.

Com relação aos metadados administrativos técnicos, o Dataverse usa JOHVE para verificar e identificar os formatos dos arquivos. O Dataverse registra o formato dos arquivos em metadados. Para dados tabulares, o Dataverse adota o Universal Numerical Fingerprint (UNF) (ALTMAN; KING, 2007), que é uma assinatura única do objeto digital, em que o valor *hash* é obtido a partir de uma representação canônica, permitindo que um mesmo objeto, armazenado em várias representações (SPSS e Stata) tenha o mesmo UNF.

O Dataverse permite a definição de vocabulários controlados para campos, cadastrados por meio de tabela. Com relação à documentação dos dados, o sistema permite que um arquivo de documentação seja armazenado no *dataset* receba a etiqueta "Documentação". Informações de *Codebook*/Livro de Códigos, isto é, sobre as variáveis de um arquivo de dados, são representadas em metadados, que são extraídos dos conjuntos de dados tabulares submetidos.

Em relação a produção dos conjuntos de dados (*datasets*), a figura 3 apresenta os elementos de DSpace e Dataverse para submissão de item e para determinar papéis relativos à publicação de um *dataset*, respectivamente.

Figura 2 – Fluxo de submissão de DSpace e papeis de Dataverse



a) Fluxo de Submissão de DSPace

b) Papéis de Dataverse

Fonte: Elaborado pelos autores (2021)

O DSpace permite, em sua instalação padrão, quatro etapas de fluxo, que podem ser usadas para a definição de um fluxo de submissão para cada coleção (figura 2a). Por meio desse fluxo, usuários de grupos autorizados e devidamente autenticados podem submeter itens à coleção. A submissão envolve a construção do pacote de submissão, com o preenchimento dos metadados e a carga dos arquivos. O fluxo também pode determinar quais grupos de usuários serão responsáveis por aprovar ou rejeitar um item, e quais grupos irão conferir e corrigir os metadados preenchidos pelos produtores. Ao submeter um item, o produtor assinala que concorda com os termos de submissão, que é então adicionado ao pacote. Por meio do fluxo de submissão, da autorização e da autenticação dos usuários, da concordância com o termo de submissão, do armazenamento do termo e do registro de ações em metadados de proveniência, o DSpace permite a transferência de custódia do objeto submetido ao repositório.

O DSpace também permite a submissão por máquinas, via rede e/ou em lote, desde que suas estruturas estejam de acordo com as suportadas pelo *DSpace Archival Information Package e DSpace* METS SIP. Por meio do protocolo Sword, outros sistemas computacionais podem submeter itens ao repositório. Coleções podem ser alimentadas a partir de metadados/itens que estão em outros repositórios, via protocolos Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) e Open Archives

Initiative Object Reuse and Exchange (OAI-ORE), funcionando como agregador de informações em um ambiente federado de repositórios.

No Dataverse, para submissão e aprovação de um *dataset*, políticas e estratégias de submissão são construídas por meio da criação e da atribuição de papéis a grupos específicos de usuários (figura 2b). A transferência da custódia dos dados ao repositório ocorre por meio dos recursos para incluir e publicar *datasets*. Produtores são devidamente autenticados e usuários com papéis devidamente atribuídos têm autorizações de submeter e de publicar conjuntos de dados. O ambiente registra as versões, com data e pessoa que publicou.

O ambiente possibilita que um *dataset* criado seja submetido para revisão antes de ser publicado, notificando os curadores, que decidem então se o *dataset* será publicado ou se retornará aos autores. Sendo publicado, o dataset assume a Versão 1, e a informação de quem a publicou é registrada. Quando um *dataset* é atualizado, pela edição de metadados ou remoção/adição de arquivos, uma nova versão *draft* do *dataset* é criada, que, se publicada, torna-se a nova versão do *dataset*. Cada usuário tem acesso às notificações a ele direcionadas, como, por exemplo, a uma notificação de solicitação de revisão de um *dataset* no qual exerce o papel de curador. Dessa forma, Dataverse dispõe de informações para prover transparência e feedback aos envolvidos em um processo de submissão, permitindo rastreabilidade e auditoria.

O software disponibiliza uma API que permite o gerenciamento (submissão, criação, edição, remoção) de dataverses, *datasets* e arquivos, assim como de usuários, papéis e grupos. O Dataverse possibilita ainda que coleções sejam alimentadas a partir da colheita de metadados de dados armazenados em outros repositórios, via protocolos OAI-PMH. Isso permite o seu uso como agregador de informações em um ambiente federado de repositórios.

Em relação a preservação digital, é importante destacar a importância de representar itens em pacotes, independentes do software. Dessa forma, a preservação a longo prazo de um item não fica dependente da existência do software que constrói e gerencia esse item. O DSpace representa itens de forma dependente do software, com o metadados sendo armazenados na base de dados de DSpace e os arquivos em sistema de arquivos.

Entretanto, um importante recurso de DSpace para apoio à preservação digital é a funcionalidade que permite a exportação/importação dos pacotes de armazenamento em

formato no padrão METS, com a descrição da estrutura desse pacote. Isso permite que ações de preservação digital sejam realizadas por aplicativos externos ao DSpace. Por exemplo, o ambiente de preservação Archivematica, que pode extrair os pacotes de DSpace e passa a atuar na preservação a longo prazo desses pacotes.

O DSpace permite a participação do repositório em redes cooperadas de preservação digital que usam o protocolo Lockss[23], em que os dados são replicados nos repositórios membros, possibilitando a recuperação em caso de perda. A Rede Cariniana[24] é uma iniciativa brasileira, promovida pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), da qual fazem parte repositórios nacionais que usam DSpace.

DSpace tem microsserviços de curadoria digital que permitem verificar a integridades dos *links* dos objetos, checar a ocorrência de vírus e checar se campos obrigatórios estão presentes (DSPACE 6.x DOCUMENTATION, 2018). Possui o recurso *Curation Task*, que permite o desenvolvimento de novos microserviços (via extensão de classes do software) e que podem ser chamadas no *workflow* de submissão. Por fim, quando um item é removido, ele ainda fica disponível aos administradores gerais do DSpace, podendo ser restaurado ou removido em caráter permanente. Repositórios digitais que usam DSpace obtiveram certificações Data Seal Approval[25] ou Core Trust Seal[26], como Dryad[27] e Datashare[28], demonstrando que a ferramenta não traz impedimentos para tal.

O Dataverse permite que metadados e arquivos que compõem *datasets* sejam extraídos do ambiente via interface de programação (API). Isso possibilita que ações de preservação digital sejam realizadas por aplicativos externos ao Dataverse. Esta API pode ser configurada para integração com o ambiente Archivematica[29], que assume a responsabilidade de preservar os *datasets* a longo prazo, enquanto o Dataverse assume a função de dar acesso aos dados.

---

[23] Lockss - https://www.lockss.org/about/why-lockss
[24] Rede Cariniana - http://cariniana.ibict.br/
[25] Repositórios de Dados em DSpace com Certificação DSA, cadastrados no Diretório Re3data - https://www.re3data.org/search?query=&software%5B%5D=DSpace&certificates%5B%5D=DSA
[26] Repositórios de Dados em DSpace com Certificação Core Trust Seal, cadastrados em Re3data - https://www.re3data.org/search?query=&software%5B%5D=DSpace&certificates%5B%5D=CoreTrustSeal
[27] Dryad - https://datadryad.org
[28] Datashare - https://datashare.is.ed.ac.uk/
[29] Archivematica – Dataverse - https://wiki.archivematica.org/Dataverse

O Dataverse verifica a integridade dos arquivos quando esses são submetidos, checando seus formatos e extraindo metadados para controle de fixidez (*hash code*). Para dados tabulares, além de checar formatos, o Dataverse extrai as variáveis, gera uma representação dos dados em formato canônico (.tab), e extrai e armazena o UNF, número *hash* criptografado (assinatura), que pode ser usado para identificar unicamente uma versão de conjunto de dados.

O Dataverse permite que dados sejam removidos. Os *datasets* removidos não são excluídos fisicamente. Repositórios digitais que usam Dataverse obtiveram certificações Data Seal Approval ou Core Trust Seal como por exemplo, Odum Institute Archive Dataverse[30], demonstrando que a ferramenta não traz impedimentos para tal.

Nas questões de uso dos conjuntos de dados, o DSpace permite a navegação por expressões de busca e por facetas. A navegação é configurada com a indicação do metadados que serão disponibilizados. Na busca, o software ordena resultados por estimativa de importância.

Objetos e metadados são acessados via serviços de identificadores globais e via protocolos abertos, pois são usados os serviços do Digital Object Identifier (DOI) ou Handle System, para identificação, e HTTP e HTTPS, para acesso e transferência de informações entre o cliente e o repositório. O DSpace foi concebido no serviço Handle System e o usa para prover identificadores globais e persistentes para seus itens. Ele permite que identificadores DOI sejam registrados em DataCite[31] e EZID[32], por meio de funções que chamam as APIs de registros dessas agências (DSPACE 6.x DOCUMENTATION, 2018).

Em relação às estatísticas básicas de uso, DSpace fornece frequências de itens arquivados, visualizações de *bitstream*, coleções e comunidades, *logins* de usuários, pesquisas realizadas e Requisições OAI-PMH. Uma série de outras estatísticas específicas são registradas. Por exemplo, o *software* Solr fornece informações sobre a quantidade de visitas às homepages de comunidades, às coleções e aos itens, bem como sobre os dez países e as dez cidades com mais visitantes.

O Dataverse permite busca básica e avançada e navegação por facetas. A busca básica recupera informação nos conteúdos, incluindo dataverses, *datasets* e arquivos. Os

[30] Repositório ODUM - https://dataverse.unc.edu/dataverse/odum
[31] DataCite – Atribuindo DOI - https://www.datacite.org/dois.html
[32] EZID – DOI - https://ezid.cdlib.org/

resultados podem ser ordenados por relevância, por nome e por data. A exploração dos dados também é apoiada por facetas, que são definidas pelo criador do dataverse, a partir de metadados, como ano de publicação, categoria do dataverse, assunto, afiliação do autor, tipo de arquivo etc. Cada entidade dataverse pode ser configurado para exibir um conjunto específico de facetas.

Por meio do componente Termos, de um *dataset*, o Dataverse disponibiliza aos usuários licenças de uso e informações de como ter acesso a dados que são de acesso restrito. O Dataverse disponibiliza informações de proveniência presentes nos metadados de citação (como autor, depositante, data). O software também apresenta informações de proveniência relevante às versões do *dataset* (data e publicador), assim como metadados de proveniência associados a arquivos.

O Dataverse dá um grande destaque à citação. Os metadados obrigatórios são compatíveis com o padrão de DataCite, permitindo a integração com serviços de citação, e a geração de informação de citação a ser disponibilizada ao usuário. O padrão de citação usado no Dataverse oferece tanto o reconhecimento adequado dos autores, como o uso de identificadores globais persistentes. O software usa impressões digitais (UNF) para permitir a verificação da autenticidade dos dados citados. O ambiente também permite a exportação da citação em formatos EndNote, RIS e BibTeX.

No Dataverse, objetos e metadados são acessados via serviços de identificadores globais e via protocolos abertos: permite o uso dos serviços DOI e Handle System para identificação, e HTTP e HTTPS, para acesso e transferência de informações entre o cliente e o repositório. O uso de DOI implica registrar o identificador em agências de registro, como DataCite[33], por meio de funções que chamam as APIs de registros dessas agências. O Dataverse possibilita que o identificador ORCID seja representado junto aos nomes de pessoas, como depositantes e autores de *datasets*.

Datasets e dataverses, quando publicados, são de acesso público. Arquivos publicados podem ter acesso restrito, mas nesse caso, o acesso só é possível por usuários autenticados e autorizados. O Dataverse dispõe do recurso "Livro de Visitas" (*Guestbook*) para estabelecer um relacionamento com os usuários dos dados. Por meio desse recurso, um *dataset* pode ser configurado para solicitar que o usuário, ao baixar um arquivo, responda a um questionário. O ambiente permite a criação de vários questionários, a

[33] DataCite – Atribuindo DOI - https://www.datacite.org/dois.html

serem usados como Livros de Visita de vários *datasets*, contendo questões de texto livre e múltipla escolha.

O Dataverse apresenta o número de downloads para cada dataverse, *dataset* ou arquivo e disponibiliza a API de Métricas (dataverses, *datasets* e arquivos adicionados por mês, *dowloads*, dataverses por categorias, *datasets* por assunto). Com essa API, é possível construir scripts que rodam em navegadores e que exibem métricas, como no caso do Repositório Harvard Dataverse[34] .

Quanto ao embargo, este é gerenciado por meio do módulo Termos, em que o *dataset* passa a ser de acesso restrito. Para dados embargados, informações de como obter acesso a esse objeto são apresentadas aos usuários. O Dataverse não gerencia período de embargo, embora este recurso esteja prometido na documentação para versões futuras.

Os softwares estudados oferecem uma solução integrada, isto é, suas funcionalidades abrangem (em maior ou menor grau) as entidades funcionais do modelo OAIS: submissão, armazenamento a longo prazo, acesso, gestão de metadados, administração e preservação digital. A vantagem da solução integrada é que tudo está presente em um único software, trazendo facilidades para instalação e operação conjunta, e proporcionando robustez do ambiente. A desvantagem está na dificuldade em adaptar o software às características do repositório, isto é, para estender o software para que este passe a atender a aspectos do ambiente do repositório não contemplados pelo software original.

O quadro 3 resume comparativamente as principais características de DSpace e Dataverse.

Quadro 3 – DSpace e Dataverse

Desenvolvimento, Manutenção e Uso do Software
• Ambos são software livre, com esquema de distribuição, atualizações e versionamento.
Representação do Ambiente do Repositório
• Ambos permitem a representação de comunidades ou grupos e subdivisões hierárquicas destes. • Ambos permitem que grupos/comunidades possuam políticas próprias de submissão e acesso • Nenhum foi desenvolvido com ênfase na interoperabilidade CRIS.
Representação dos Conjuntos de Dados
• Dataverse foi planejado para representar conjuntos de dados. Já DSpace foi planejado para representar coleções de itens, em que um conjunto de dados pode ser representado como um item.

[34] Métricas de Harvard Dataverse - https://dataverse.org/metrics

• Ambos podem armazenar conjuntos de dados compostos por múltiplos arquivos, em vários formatos, com metadados e termos de uso/licença. No Dspace, os arquivos de um item podem ser organizados em diretórios de um nível, e a licença é um documento armazenado no item. • Dataverse usa aplicativo para verificar os formatos dos arquivos (JHOVE) e possui recursos especiais para tratar arquivos tabulares e geoespaciais. DSpace permite a definição de política de formatos (formatos aceitos) e identifica o formato pela terminação do arquivo. • Dataverse possui módulo nativo para gerenciar versões, que são atreladas ao processo de publicação do *dataset*, com a geração de citações considerando essas versões. DSpace possui um módulo de versão que foi incorporado na versão do 3 do software. • Ambos armazenam informações do pacote (metadados, estrutura) em banco de dados e os arquivos, em sistemas de arquivos. DSpace exporta item na forma de pacote de armazenamento, com estrutura descrita através do padrão METS. Dataverse permite a extração das informações do conjunto de dados via API.
Descrição e Documentação dos Conjuntos de Dados
• Ambos permitem a definição de esquemas de metadados. • DSpace dispõe do esquema Dublin Core pré-configurado. Dataverse possui um esquema de citação obrigatório. • Dataverse dispõe de esquemas pré-configurados para as Ciências Sociais e Humanidades, Astronomia e Astrofísica e Ciências da Vida e apresenta mapeamentos destes esquemas para esquemas padrões representados em XML, como Dublin Core, DDI, Datasite, VOResource, ISA-Tab. DSpace disponibiliza recurso (regras XSLT) para mapear esquemas de metadados para representações em esquemas padrões XML, já dispondo de vários mapeamentos. • Em Dspace, informações de proveniência referentes à submissão (metadados administrativos) são registrados em metadado. Dataverse registra e apresenta ao usuário as versões de um *dataset*, seus publicadores, data de publicação e mudanças ocorridas, que são informações que documentam ações administrativas referentes à proveniência. O esquema de citação de dataverse permite a representação informações que dizem respeito à proveniência (autor, produtor, colaborador, depositante, fontes etc.). • Ambos registram as informações sobre as estruturas dos itens e dos *datasets* em banco de dados, respectivamente. DSpace exporta um item na forma de pacote de armazenamento, utilizando o padrão METS para metadados estruturais. • O Dataverse permite a definição de vocabulários controlados para campos, cadastrados por meio de tabela. DSpace permite o controle de vocabulário a partir de recursos de configuração da interface. • Documentos do tipo Codebook podem ser armazenados como arquivos em DSpace e Dataverse. Dataverse permite que esses arquivos possam ser etiquetados como tal. Dataverse usa o esquema de metadados DDI Codebook.
Produção dos Conjuntos de Dados
• Ambos peritem a especificação de processos de submissão, que são realizados por usuários autenticados e autorizados. Registram as ações e permitem que os envolvidos sejam informados sobre a submissão. Em DSpace, o processo de submissão é especificado pela configuração de um fluxo. Em Dataverse, o processo é estabelecido através da atribuição de papéis. • Ambos permitem a submissão por máquina, via Sword e API. • Ambos realizam a autenticação de usuário, e permitem que essa autenticação possa ser feita por ambientes externos. DSpace, via LDAP, Shibboleth, IP Address e certificado X.509. Dataverse, via Shibboleth e OAuth2. • Ambos possuem uma interface de usuário que conduzem a construção e a ingestão do material submetido, proporcionando conformidade com as estruturas especificadas para item e *dataset*. Ambos apresentam etapas em que usuários autorizados podem ser incumbidos de verificar e validar o material de forma manual. Extraem checksum dos arquivos. Dataverse verifica automaticamente o formato e possui microserviços que analisam dados tabulares e geoespaciais. • Ambos permitem a definição de licenças/ termos de uso. Em DSpace, o produtor acorda com uma licença de uso, que é armazenada em como um arquivo do item. Em Dataverse, os termos de uso são preenchidos. • Ambos permitem que descrições de conjuntos de dados possam ser produzidas a partir de colheita, via OAI-PMH.

Armazenamento a Longo Prazo e Planejamento da Preservação
• Ambos registram estrutura e metadados em banco de dados e os arquivos, em sistemas de arquivos. DSpace permite a exportação de pacote de armazenamento em conformidade com requisitos e padrões de preservação digital. Dataverse dispõe de uma API que permite a extração das informações de um *dataset*. • Ambos permitem que seus conjuntos sejam extraídos por ambientes externos que podem ser acoplados para tratar da preservação a longo prazo, como Archivematica. DSpace possui integração com Lockss. • DSpace registra e permite o cadastramento dos formatos registrados aceitos, dispõe de microserviços para verificação de vírus, links e metadados obrigatórios. Dataverse verifica integridade dos formatos dos arquivos. • Repositórios digitais que usam tanto Dataverse quanto DSpace obtiveram certificações Data Seal Approval ou Core Trust Seal.
Acesso e Uso dos Conjuntos de Dados
• Ambos permitem a busca por expressões e navegação por facetas. • Ambos disponibilizam aos usuários os conjuntos de dados, seus metadados e termos/licença, que podem ser obtidos via os protocolos abertos HTTP e HTTPS. Dataverse apresenta informações de citação. • Ambos usam identificadores globais persistente (Handle e DOI) para identificar conjuntos de dados. • Ambos permitem que metadados sejam colhidos via OAI-PMH. DSpace permite que itens possam ser obtidos via OAI-ORE. Em Dataverse, informações do *dataset* podem ser extraídas via API. • DSpace permite políticas de restrição de acesso a comunidades, coleções e itens. Registra e gerencia embargos. Dataverses e *datasets*, quando publicados, são de acesso público. Arquivos de *datasets* publicados podem ter acesso restrito, por usuários autenticados e autorizados. • Dataverse possui integração com ferramentas para exploração de dados espaciais (WordMap) e tabulares (TwoRaves). DSpace oferece uma interface configurável para integração com visualizadores de objetos. • Ambos apresentam estatísticas de uso. Dataverse dispõe do recurso “Livro de Visitas” (Guestbook) para estabelecer um relacionamento com os usuários dos dados.

Fonte: Elaborado pelos autores (2021).

No Quadro 3 podemos observar muitas similaridades entre os dois softwares. Atualmente, vários softwares livres são usados para repositórios de dados da pesquisa. Alguns foram desenvolvidos especificamente para dados de pesquisa, como o Dataverse, outros foram originalmente desenvolvidos para diferentes propósitos, como para repositório institucional de documentos, como o DSpace. Nesse ponto, o Dataverse diferencia-se quando a questão é o armazenamento de dados.

4 CONSIDERAÇÕES FINAIS

O Dataverse é um software integrado para publicação, compartilhamento e armazenamento de dados. Ele traz facilidades para representar cenários que são compostos por diversas entidades hierárquicas (como universidades, unidades ou grupos), que são autônomas, isto é, que têm poder para definir quem pode criar, autorizar

a publicação ou acessar conjuntos de dados, estabelecer licenças e definir se o uso dos dados somente deverá ser feito mediante solicitação. Ele também permite a configuração e o uso de esquemas de metadados (compatíveis com DDI Lite, DDI Codebbok, DUblin Core, DataCite, VORResource, ISA-Tab), gerencia versões de conjuntos de dados, identifica univocamente conjuntos de dados (considerando versões) de forma universal e persistente (sistemas DOI ou Handle System), disponibiliza metadados de citação e uma estrutura para citação que envolve a verificação da fixidez do material citado. Permite o armazenamento de documentos complementares junto a conjunto de dados, a adição de ferramentas de análise de dados, a customização de interfaces, o uso de serviços de caracterização de formatos, a submissão por máquinas (Sword) e a colheita de metadados (OAI-PMH). É usado por instituições (heiData[35]/Univ. Heidelberg), por grupos de instituições (DataverseNL/Holanda, Abacus[36]/Canada, Texas Digital Library/EUA) e para repositórios temáticos (ICRIAT[37]/Agricultura) e multidisciplinares (Australian Data Archive[38]). Alguns repositórios que usam o Dataverse são repositórios digitais confiáveis certificados, como o Australian Data Archive e o TiU Dataverse[39]/DatavereNL, demonstrando que o software acolhe a necessidades desse tipo. O Dataverse também atende a requisitos FAIR, conforme atestam Wilkinson e outros (2016).

O DSpace é um software desenvolvido para repositório institucional. Assim como o Dataverse, permite a representação de unidades e de subunidades autônomas, com a configuração de fluxos específicos de submissão; o uso de diversos esquemas de metadados; a identificação universal e persistente por meio de Handle System; a distribuição de cópias de segurança (Lockss); a submissão por máquina (Sword) e a colheita de metadados (OAI-PMH). O DSpace é usado por reconhecidos repositórios de dados, como o repositório multidisciplinar Dryad[40] e o repositório institucional DataShare[41]. O DataShare[42] é um exemplo de repositório confiável e certificado, demonstrando que repositórios de dados em DSpace podem ser certificados como confiáveis.

[35] heiDATA. https://heidata.uni-heidelberg.de/
[36] Abacus Dataverse Network https://abacus.library.ubc.ca
[37] Int. Crops Research Institute for the Semi-Arid Tropics- http://dataverse.icrisat.org/
[38] Australian Data Archive. https://dataverse.ada.edu.au/
[39] TiU - CoreTrustSeal https://www.coretrustseal.org/wp-content/uploads/2018/04/Tilburg-University-Dataverse.pdf
[40] Dryad - https://datadryad.org/
[41] Edinburgh DataShare. https://datashare.is.ed.ac.uk/
[42] Datashare - Data Seal Approval https://assessment.datasealofapproval.org/assessment_175/seal/pdf/

Como o Dataverse foi desenvolvido para repositório de dados, a representação e a gestão automatizada dos conjuntos de dados são estruturadas por meio do conceito *dataset*, que inclui dados, metadados de citação, metadados específicos, documentação adicional, citação, gerenciamento de versões etc. Já o DSpace está estruturado no conceito de coleção de itens, com cada item sendo composto por pastas (*bundles*) que contêm arquivos (*bitstreams*). No caso do uso de DSpace para gerenciar dados, pode ser necessário configurar metadados, fluxos e interfaces de usuário para conduzir a submissão de dados, como feito no repositório DataShare.

Considerando os dois softwares, conclui-se que o Dataverse possui recursos para configuração de vários tipos de ambientes de repositório, incluindo hierarquias organizacionais, políticas de gestão distintas para unidades ou grupos, incluindo esquemas de metadados e licenças. Isso é possível no DSpace, entretanto exige uma série de adaptações nas configurações.

REFERÊNCIAS

ALBAGLI, Sarita; CLINIO, Anne; RAYCHTOCK, Sabryna. Ciência Aberta: correntes interpretativas e tipos de ação. **Liinc em Revista**, Rio de Janeiro, v. 10, n. 2, p. 434-450, novembro 2014. Disponível em: http://revista.ibict.br/liinc/article/view/3593/3072 . Acesso em: 17 abr. 2020.

ALTMAN, M.; KING, G. A Proposed Standard for the Scholarly Citation of Quantitative Data. **D-Lib Magazine**, 13, 2007. Disponível em: http://www.dlib.org/dlib/march07/altman/03altman.html Acesso em: 17 abr. 2020.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. Producer-Archive Interface Methodology Abstract Standard. CCSDS 651.0-M-1. 2004 https://public.ccsds.org/Pubs/651x0m1.pdf

CROSAS, M. The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. **DLIB Magazine**, 2011. Disponível em: http://dlib.org/dlib/january11/crosas/01crosas.html . Acesso em: 17 abr. 2020.

DATA CITATION SYNTHESIS GROUP. **Joint Declaration of Data Citation Principles**. San Diego: FORCE11, 2014. Disponível em: https://doi.org/10.25490/a97f-egyk . Acesso em: 17 abr. 2020.

DSPACE 6.x DOCUMENTATION. 2018. Disponível em: https://wiki.duraspace.org/display/DSDOC6x/Item+Level+Versioning . Acesso em: 17 abr. 2020.

DURASPACE. **DSpace**. Disponível em: https://duraspace.org/dspace/download/. Acesso em: 17 abr. 2020.

GABRIEL JUNIOR, Rene Faustino; ROCHA, Rafael Port da; CAREGNATO, Sônia Elisa; PAVÃO, Caterina Marta Groposo; PASSOS, Paula Caroline Schifino Jardim; BORGES, Eduardo Nunes;

VANZ, Samile Andréa de Souza; AZAMBUJA, Luís Alberto Barbosa. Acesso aberto a dados de pesquisa no Brasil: mapeamento de repositórios, práticas e percepções dos pesquisadores e tecnologias. **Ciência da Informação**, Brasília, v. 48, n. 3, 2019. Disponível em: http://revista.ibict.br/ciinf/article/view/4958 . Acesso em: 17 abr. 2020.

HENNING, Patricia Correa et al. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, Porto Alegre, v. 25, n. 2, maio/ago. 2019. Disponível em: https://www.seer.ufrgs.br/EmQuestao/article/view/84753 . Acesso em 09 jun. 2021.

INTERNATIONAL STANDARD ORGANIZATION. **ISO 16363:** Space data and information transfer systems — Audit and certification of trustworthy digital repositories. 2012. Disponível em: https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/iso16363. Acesso em: 17 abr. 2020.

LEEUW, L. Data Seal of Approval (DSA). Data Archiving and Networked Services. 2019. Disponível em: https://doi.org/10.17026/dans-28z-njxq . Acesso em: 17 abr. 2020.

ROCHA, R. P.; *et al.* Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas: relatório. 2018. Disponível em: http://hdl. handle.net/10183/185126.

SAYÃO, Luis Fernando; SALES, Luana Farias. Curadoria Digital: um novo patamar para preservação de dados digitais de pesquisa**. Informação & Sociedade Estudos**, João Pessoa, v. 22, n. 3, p. 179-191, set./dez. 2012.

SAYÃO, Luis Fernando; SALES, Luana Farias. **Guia de Gestão de dados de pesquisa para bibliotecários e pesquisadores.** Rio de Janeiro: CNEN, 2015. 90 p. Disponível em: http://www.cnen.gov.br/component/content/article/75-cin/material-didatico-cnen/160-guia-de-gestao-de-dados-de-pesquisa . Acesso em: 22 maio. 2019.

SAYÃO, Luis Fernando; SALES, Luana Farias. Subsídios para a construção de um modelo de avaliação de sistemas de gestão de dados de pesquisa. **PontodeAcesso**, Salvador, v. 12, n. 3, p. 80-108, dez. 2018.

THE CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. Recommendation for Space Data System Practices: Reference Model for an Open Archival Information System (OAIS). Recommended Practice, Issue 2, June 2012, Washington. Disponível em: https://public.ccsds.org/pubs/650x0m2.pdf . Acesso em: 17 abr. 2020.

WILKINSON, M. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. ***Sci Data,*** n. 3, 2016. Disponível em: https://doi.org/10.1038/sdata.2016.18 . Acesso em: 17 abr. 2020.

Recebido em: 23 de março de 2021
Aprovado em: 12 de junho de 2021
Publicado em: 21 de junho de 2021