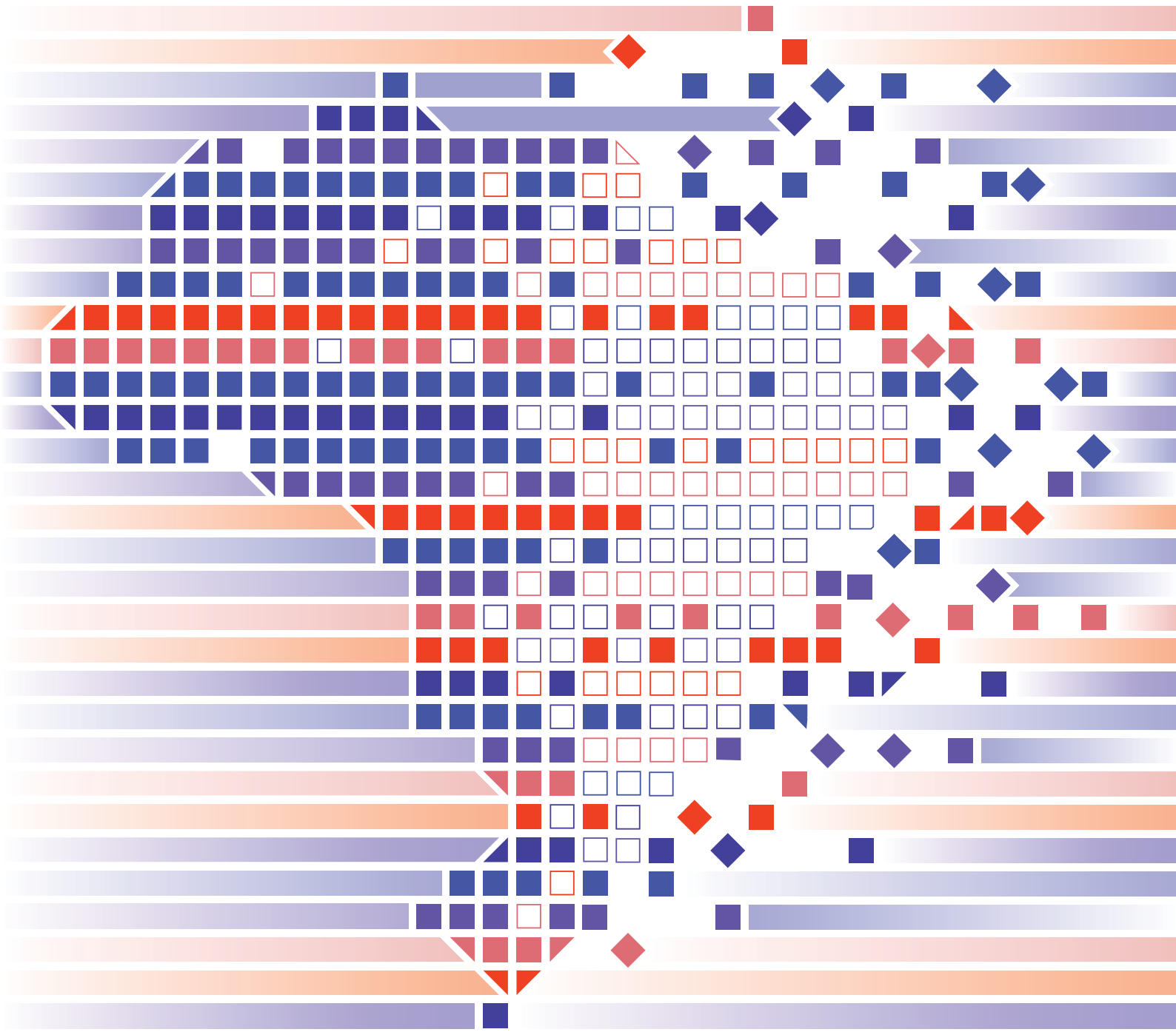
ACESSO ABERTO A DADOS DE PESQUISA NO BRASIL

Soluções Tecnológicas

RELATÓRIO – 2018



RDP BRASIL
Rede de Dados de Pesquisa

Coordenador
Rafael Port da Rocha (Universidade Federal do Rio Grande do Sul)

Coordenadora adjunta
Sônia Elisa Caregnato (Universidade Federal do Rio Grande do Sul)

Pesquisadores
Universidade Federal do Rio Grande do Sul
Caterina Groposo Pavão
Paula Caroline Schifino Jardim Passos
Rene Faustino Gabriel Junior
Samile Andréa de Souza Vanz
Universidade Federal de Rio Grande
Eduardo Nunes Borges
Luís Alberto Barbosa Azambuja

Bolsista
Universidade Federal do Rio Grande do Sul
Iván Andrés Fornos Angues
Victor Andrews Garcia Lima

Contato
e-mail: dadosdepesquisa@rnp.br
site: https//dadosdepesquisa.rnp.br

Rua Ramiro Barcelos, 2705 - Campus Saúde
Sala: 106 - Anexo 1
Brasil - Porto Alegre - RS - CEP 90.035-007
Telefone: +55(51)3308.5942

Este relatório é licenciado sobre a licença CC BY - Creative Commons Attribution 4.0 International License.

Como citar:
ROCHA, Rafael Port da; AZAMBUJA, Luís Alberto Barbosa; BORGES, Nunes Borges; GABRIEL JUNIOR, Rene Faustino; CAREGNATO, Sônia Elisa; PAVÃO, Caterina Groposo; VANZ, Samile Andrea de Souza; PASSOS, Paula Caroline Schifino Jardim. Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas: relatório 2018. Disponível em: <http://hdl.handle.net/10183/185126>.

R672 ROCHA, Rafael Port da

Acesso aberto a dados de pesquisa no Brasil : soluções tecnológicas : relatório 2018 / Rafael Port da Rocha et al. - Porto Alegre, RS : UFRGS, 2018. 75p.

Relatório do projeto RDP Brasil: Rede de Dados de Pesquisa Brasileira, – Universidade Federal do Rio Grande do Sul; Universidade Federal do Rio Grande.

1. Dados abertos de pesquisa. 2. Compartilhamento de dados. 3. Tecnologias. I. Título

Resumo

Este relatório apresenta os resultados do projeto de pesquisa **Rede de Dados de Pesquisa Brasileira (RDP Brasil)**, desenvolvido pela Universidade Federal do Rio Grande do Sul (UFRGS) em parceria com a Universidade Federal do Rio Grande (FURG), sob coordenação executiva da Rede Nacional de Ensino e Pesquisa (RNP) e do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Os resultados estão relacionados ao quarto objetivo do projeto, que identifica soluções tecnológicas para a construção de repositório para Acesso Aberto a Dados de Pesquisa (AADP). 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. Com base no modelo OAIS, foram elaborados 56 critérios para analisar as soluções tecnológicas classificados em 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. Adicionalmente, abordou-se o software CKAN. Considerando as três soluções, 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 com adaptações por ter sido desenvolvido para repositórios institucionais e estar estruturado a partir do conceito de coleção de itens. Por sua vez, o CKAN, apesar de possuir menos funcionalidades, é uma alternativa quando usado como serviço de publicação e acesso, com a submissão e preservação digital sendo realizados por outros ambientes.

Palavras-chave:
Dados abertos de pesquisa. Compartilhamento de dados. Tecnologias.

Sumário

1 Introdução

A Rede Nacional de Ensino e Pesquisa (RNP) lançou, em 2017, um edital para promover pesquisas sobre compartilhamento dos dados coletados, gerados e utilizados pelos pesquisadores brasileiros. O projeto contemplado neste edital é fruto da parceria entre a Universidade Federal do Rio Grande do Sul (UFRGS) e a Universidade Federal do Rio Grande (FURG), envolvendo o Centro de Documentação e Acervo Digital de Pesquisa (CEDAP), órgão auxiliar da Faculdade de Biblioteconomia e Comunicação (FABICO), e o Centro de Processamento de Dados (CPD), ambos da UFRGS, e o Grupo de Pesquisa em Gerenciamento de Informações do Centro de Ciências Computacionais (C3), da FURG. O projeto RDP Brasil - Rede de Dados de Pesquisa Brasileira tem cinco objetivos planejados pela equipe, que são os seguintes:

1) Identificação de práticas de AADP em instituições brasileiras;
2) Mapeamento de potenciais usuários nacionais de serviços de AADP;
3) Elaboração do portal web para a comunidade nacional em AADP;
4) Levantamento comparativo dos serviços e soluções tecnológicas para compartilhamento de dados;
5) Desenvolvimento de uma proposta de solução tecnológica.

Este relatório apresenta resultados do quarto objetivo: "Levantamento comparativo dos serviços e soluções tecnológicas para compartilhamento de dados". Para identificar estes serviços e soluções, partiu-se dos modelos teóricos já conhecidos pela comunidade, com destaque para o modelo OAIS, e de diretórios internacionais de serviços para compartilhamento de dados de pesquisa, como o Re3Data.org.

O restante do texto está organizado da seguinte forma. A seção 2 apresenta uma série de condições de escopo para o levantamento das soluções tecnológicas. Na seção 3 são detalhados os critérios desenvolvidos para análise das soluções tecnológicas, organizados em categorias e relacionados com os princípios FAIR. A seção 4 detalha cada solução tecnológica estudada e as caracteriza em relação aos critérios definidos. Por fim, a seção 5 apresenta as considerações finais.

2 Considerações de Escopo para Análise das Soluções Tecnológicas

Este estudo investiga soluções tecnológicas para desenvolvimento de repositório de dados de pesquisa no cenário nacional. Em função de uma grande diversidade de tipos de repositórios, as seguintes considerações são apresentadas com relação ao escopo da investigação.

Repositórios de Dados de Pesquisa podem ser **institucionais** (como heiDATA[1], data.bris[2], DataShare[3]), **multidisciplinares** (como Figshare[4], dataHub[5], Zenodo[6]), **disciplinares** (como ICPSR[7], UK Data Archive[8], Protein Data Bank[9]) e **nacionais** (como Dataverse NL[10] e Easy[11]). Repositórios **institucionais**, **nacionais** ou **multidisciplinares** armazenam uma grande diversidade de tipos de dados, abrangendo dados com características das mais diversas áreas de pesquisa, sendo assim caracterizados como repositórios de **cauda longa** dos dados. Repositórios **disciplinares** armazenam dados que seguem padrões e recomendações de disciplinas específicas. Muitas dessas áreas possuem padrões e soluções tecnológicas específicas, que são coordenadas por associações ou organizações, como Elixir[12], da área das Ciências da Vida.

Quanto ao tipo de propósito, esse estudo adota duas abordagens. Para repositórios institucionais, multidisciplinares e nacionais, cujos dados são de uma grande diversidade de tipos (cauda longa), esse estudo investiga soluções tecnológicas disponíveis no mercado (*software* de prateleira), preferencialmente na forma de *software* livre, não envolvendo o estudo de soluções desenvolvidas especificamente para um determinado repositório. Isso deve-se ao fato que soluções de prateleira minimizam custos, sendo a solução mais comum para esses tipos de repositório. Para repositórios disciplinares, estuda uma arquitetura desenvolvida pelo projeto EUDAT[13] para prover infraestrutura de repositório.

Repositórios podem adotar vários modelos de funcionamento, quando consideramos suas relações com os produtores e os custodiadores da informação. Um único repositório pode atuar como repositório de dados institucional de várias instituições, como Dataverse NL[10] e Texas Data Repository (TDL)[14]. Nesse caso, a participação de cada instituição é determinada por acordos, como no modelo de negócio do repositório TDL[15].

1 Repositório de dados da Universidade de Heidelberg – HeiData (Dataverse) - https://heidata.uni-heidelberg.de/
2 Repositório de dados da Universidade de Bristol (KCAN)- https://data.bris.ac.uk/data/
3 Repositório de dados da Universidade de Edinburg (DSpace) - https://datashare.is.ed.ac.uk/
4 Repositório de dados multidisciplinar Figshare DSpace) - https://figshare.com/
5 Repositório de dados Data Hub - https://datahub.io/
6 Repositório de dados Zenodo - https://zenodo.org/
7 Repositório de dados ICPSR - https://www.icpsr.umich.edu/icpsrweb/ICPSR/
8 Repositório de dados disciplinar UK Data Archive - http://data-archive.ac.uk/
9 Repositório Protein Data Bank - https://www.wwpdb.org/
10 Repositório de dados Dataverse NL - https://dataverse.nl/
11 Repositório de dados para preservação a longo prazo EASY - https://easy.dans.knaw.nl/ui/home
12 Elixir – Infraestrutura distribuída para dados de pesquisa para Ciências da Vida - https://www.elixir-europe.org/
13 Projeto EUDAT - https://eudat.eu/
14 Repositório Texas Data Repository (TDR) - https://data.tdl.org/
15 Acordos de Participação no Texas Data Repository - https://www.tdl.org/members/membership/

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. Nesse contexto, um estudo é considerado como uma coleção de arquivos de dados resultantes da coleção intencional de dados, através de solicitação, observação ou coleta de fontes secundárias, com propósito e metodologia descritos[16]. Por outro lado, um repositório pode ter como estratégia permitir que grupos de pesquisa armazenem os conjuntos de dados de seus vários estudos.

Para atender a essas estratégias de funcionamento, soluções tecnológicas devem prover recursos para representação de estudos, grupos ou organizações, assim como mecanismos para permitir a configuração de políticas distintas para submissão e gestão para grupos ou estudos. Isso envolve também estratégias para gestão e autenticação de usuários. Por exemplo, Dataverse NL[10] utiliza o serviço de autenticação federado SURFconext [17] [18].

Considerando o ciclo de vida do dado, repositórios podem ter como política permitir o armazenamento e o compartilhamento dos dados que acompanham uma pesquisa ao longo de seu andamento, outros adotam a política de armazenar somente naqueles dados de pesquisa que devem ser preservados a longo prazo. Dataverse NL[10], por exemplo, permite o armazenamento, o compartilhamento e o registro de dados de pesquisa durante o período da pesquisa, mas o serviço prescreve após dez anos do término da pesquisa[19]. Já EASY [11] atende o armazenamento a longo prazo, sendo opção de preservação digital para dados de Dataverse NL [10].

Ainda considerando o ciclo de vida do dado, conjuntos de dados podem ser atualizados, resultando uma sequência de versões, como em casos em que dados que são produzidos como séries temporais, ou em que dados que são transformados ao longo das etapas do seu ciclo de vida, ou em que dados são decorrentes de transformações resultantes de reutilizações pelo grupo ou por outros grupos.

Com relação à estratégia de funcionamento, este estudo adota o seguinte: as soluções tecnológicas são investigadas considerando um repositório que seja capaz de permitir a representação e a configuração de políticas de submissão, acesso e gestão para conjuntos de dados, estudos e grupos. Esses grupos podem ser informais ou formais (institucionais) ou assumir configurações organizacionais (organização com seus órgãos, grupos e estudos). O repositório deve ser capaz de representar tanto dados produzidos durante a pesquisa, como dados a serem preservados após o seu término. O repositório deve ser capaz de permitir o gerenciamento de versões de dados decorrentes de transformações que ocorrem no ciclo de vida do dado, tanto dados coletados e processados para serem usados na pesquisa, como dados para serem reusados.

16 Modelo conceitual para Dados de Pesquisa ligado ao Padrão de Metadados DDI - http://opendatafoundation.org/ddi/srg/Papers/DDIModel_v_4.pdfb

17 Dataverse NL – Login via SurfConext - https://dans.knaw.nl/nl/over/diensten/DataverseNL/about

18 SurfConext Serviço de Autorização e Autenticação permitindo o acesso de login único a vários serviços de nuvem - https://www.surf.nl/en/services-and-products/surfconext/what-is-surfconext/index.html

19 Dataverse NL – Políticas https://dans.knaw.nl/en/about/organisation-and-policy/information-material/Dans_factsheetDataverseENGdef.pdf

Outra questão referente à produção de dados diz respeito à distinção entre dados que são produzidos (coletados e processados) pela pesquisa, comumente chamados de "dados da pesquisa", e dados que são disponibilizados por organizações para serem usados em pesquisas, comumente chamados "dados para pesquisa". Esse estudo considera ambos os casos, pois adota a seguinte definição para dados da pesquisa:

> Registros factuais (escores numéricos, registros textuais, imagens e sons) usados como fontes primárias para pesquisa científica, e que são comumente aceitos na comunidade científica como necessários para validar os resultados da pesquisa[20].

Hoje, a implantação de um repositório é orientada por uma série de modelos de referência e princípios, como o Modelo de Referência para Repositórios, OAIS - Open Archival Information System[21] (hoje norma ISO 1472:2003[22]), e os princípios de compartilhamento FAIR[23] e de citação[24]. Além disso, para o desenvolvimento de um repositório, é importante considerar critérios para construção de repositórios digitais confiáveis, que visam a obtenção de certificações para repositórios confiáveis, como Data Seal Approval[25], Core Trust Seal[26], Nestor[27] e ISO 16363[28].

O modelo Open Archival Information System (OAIS)[22] serve como modelo de referência para o desenvolvimento de repositórios que estão comprometidos em preservar e manter acessível seus conteúdos digitais a longo prazo. É composto pelas especificações do **ambiente do repositório**, do modelo de **representação das informações** a serem preservadas e do **modelo funcional** do repositório.

O **ambiente** de um repositório é formado pelo **produtor**, pelo **consumidor** e pela **gestão.** OAIS orienta na identificação dos produtores, do processo de produção dos dados, das comunidades alvo (consumidores) e dos encarregados de sua gestão. **O modelo funcional** de OAIS especifica as entidades funcionais que compõem um repositório OAIS: **ingestão**, **acesso**, **planejamento da preservação**, **armazenamento**, **gestão de dados** e **administração**. Servem para delimitar, caracterizar e orientar os componentes de funcionamento de um repositório. O **modelo de dados** especifica e orienta a representação dos objetos digitais, incluindo pacotes de informação, representações, informações descritivas e de preservação.

OAIS também caracteriza os pacotes que representam a informação a ser submetida ao repositório (Pacote de Submissão de Informação), a informação a ser armazenada a longo prazo (Pacote de Armazenamento de Informação) pelo repositório e a

20 OECD Principles and Guidelines for Access to Research Data from Public Funding. 2007 www.oecd.org/science/sci-tech/38500813.pdf

21 Reference Model for an Open Archival Information System https://public.ccsds.org/pubs/650x0m2.pdf

22 ISO 1472:2003 https://www.iso.org/standard/57284.html

23 Princípios FAIR. https://www.force11.org/group/fairgroup/fairprinciples

24 Princípios de Citação https://www.force11.org/datacitationprinciples

25 Data Seal Approval https://www.datasealofapproval.org/

26 Core Trust Seal https://www.coretrustseal.org/

27 Nestor - http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html

28 ISO 16363 - https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/iso16363

informação a ser disseminada, isto é, entregue pelo repositório ao consumidor (Pacote de Disseminação da Informação)

Repositório confiável é aquele cuja missão é fornecer acesso de longo prazo a recursos digitais gerenciados. Um repositório com essas características aceita a responsabilidade pela manutenção a longo prazo dos recursos digitais, projeta seus sistemas de acordo com as convenções e os padrões comumente aceitos, estabelece metodologias para avaliação de sistemas que atendem às expectativas de confiabilidade da comunidade, cumpre suas responsabilidades com depositantes e usuários de forma aberta e explícita, e permite que sejam auditadas e medidas suas políticas, práticas e desempenho[29]. Critérios para verificação e certificação de repositórios digitais confiáveis medem a proximidade do repositório com o Modelo de Referência OAIS. Os principais instrumentos para certificação de repositórios digitais são Data Seal Approval [25], Trusted Core Seal [26], Nestor [27] e ISO 16363 [27].

Princípios FAIR e de citação orientam no sentido de promover o compartilhamento dos dados. Os princípios FAIR [24] indicam que dados devem ser localizáveis (*Findable*), acessíveis (*Accessible*), interoperáveis (*Interoperable*) e reusáveis (*Reusable*). Para dados serem FAIR, estes devem ser atribuídos a identificadores únicos, persistentes e globais (F). Devem ser descritos por metadados indexáveis e ricos (F), representados em linguagens formais (I), aceitos pela comunidade (R), com atributos relevantes, precisos e úteis ao contexto (R), incluindo metadados de proveniência (R) e usando vocabulários controlados que seguem princípios FAIR (I). Estes dados devem ser recuperáveis pelo seu identificador através de um protocolo de comunicação padronizado, aberto e gratuito (A). Também devem ser acompanhados de licenças claras e acessíveis (R), e referências qualificadas devem ligar (meta)dados (I).

Identificadores persistentes viabilizam a referência de longa duração aos objetos digitais, à medida que permitem que estes objetos sejam identificados de forma independente de suas localizações. Na web, identificadores persistentes evitam a perda da referência a um objeto quando ocorrer, por exemplo, a mudança do endereço do objeto. Identificadores persistentes são basicamente compostos por dois componentes: um identificador único e um serviço que localiza o objeto a longo prazo, a partir desse identificador.

DOI[30] e Handle System[31] são os principais serviços para identificadores persistentes usados em repositórios digitais. Handle System é um serviço que resolve identificadores persistentes, a partir de um registro. O serviço garante a unicidade de um identificador atribuído no contexto Handle e o roteamento das requisições de resolução[32]. É realizado através de um conjunto de servidores que cooperam para reduzir a carga da rede. DOI é um sistema que agrega metadados ao Handle System,

29 Jantz, R., Giarlo, M.: Digital Preservation - Architecture and Technology for Trusted Digital Repositories. 2006 - http://www.dlib.org/dlib/june05/jantz/06jantz.html

30 DOI - Digital Object Identifier - https://www.doi.org/

31 Handle System - https://www.handle.net/

32 Sayão, L. Interoperabilidade das bibliotecas digitais: o papel dos sistemas de identificadores persistentes - URN, PURL, DOI, Handle System, CrossRef e OpenURL. 2007 http://www.scielo.br/pdf/tinf/v19n1/06.pdf

permitindo a adição de semântica aos identificadores. A especificação DOI apresenta um conjunto nuclear de metadados, que podem ser expandidos com novos elementos para atender a domínios específicos. DOI possui um modelo de negócio em que o registro é realizado por meio de uma federação de agências. Essas agências atendem a domínios específicos, e usam esquemas de metadados específicos para atender às necessidades desses domínios. CrossRef [33] é uma agência voltada ao domínio da edição de conteúdos científicos, e DataCite[34] é uma instituição não lucrativa que provém identificadores persistentes para dados de pesquisa. Com relação aos metadados, Datacite segue o esquema DataCite Metadata Schema[35].

A citação de dados é mais ampla que a citação de publicações científicas, pois questiona referências mais granulares aos dados, incluindo subconjuntos de observações, variáveis ou outros componentes, assim como subconjuntos de um conjunto maior de dados. Essas referências granulares são frequentemente necessárias no texto para descrever o suporte probatório preciso para uma tabela de dados, figura ou análise[36]. Force 11[25] apresentou um conjunto de princípios que cobrem o propósito, a função e os atributos da citação, que são: **Importância**, **Crédito e Atribuição, Evidência**, **Identificação Única, Acesso** (aos dados, metadados e documentação), **Persistência** (identificadores e metadados que permanecem mesmo quando os dados ficam indisponíveis); **Especificidade e Verificabilidade** (informações de proveniência e fixidez suficientes para a verificação de que fatia de tempo, versão ou parte granular dos dados obtidos subsequentemente é a mesma que a que foi originalmente citada); **Interoperabilidade e Flexibilidade**.

Proveniência dos dados é enfatizada tanto nos princípios FAIR, como de citação. Pode ser definida como "informações sobre entidades, atividades e pessoas envolvidas na produção de um dado ou coisa, que podem ser usadas para formar avaliações sobre sua qualidade, confiabilidade ou fidedignidade"[37]. Isso implica em prover instrumentos para documentar como os dados foram produzidos, que envolve gerenciar as versões dos dados, identificar unicamente os conjuntos de dados (incluindo suas versões ou subconjuntos), e até mesmo registrar automaticamente o fluxo da produção dos dados, quando estes dados são produzidos a partir de sistemas de workflow[38].

Os princípios FAIR e de Citação, o modelo OAIS e os critérios de certificação de repositórios digitais confiáveis são direcionados à implementação de um repositório. Este estudo tem como enfoque investigar soluções tecnológicas para repositórios. Por isso, esse trabalho adota como estratégia analisar de que forma as soluções tecnológicas proporcionam instrumentos que permitam o desenvolvimento de repositórios de acordo

33 CrossRef - https://www.crossref.org/

34 DataCite - https://www.datacite.org/

35 DataCite Metadata Schema - https://schema.datacite.org/

36 CODATA-ICSTI . Out of Cite, Out of Mind - The Current State of Practice, Policy, and Technology for the Citation of Data. 2013 - https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf

37 Groth, P., Moreau, L.: PROV-Overview - An Overview of the PROV Family of Documents. W3C Working Group, Note 30 (2013). https://www.w3.org/TR/prov-overview/

38 Bechhofer, S., et al.: Research Objects - Towards Exchange and Reuse of Digital Knowledge. 2010 - http://dx.doi.org/10.1038/npre.2010.4626.1

com princípios FAIR, citação, proveniência, certificação e o modelo de Referência OAIS. Para identificador de objeto persistente, considera os serviços Handle System e DOI.

Outros importantes movimentos estão surgindo e envolvem também o domínio dos dados da pesquisa: Dados Abertos e Ligados e Serviços de Informação da Pesquisa.

O movimento dos Dados Abertos e Ligados refere-se a produzir e interligar dados estruturados na web. Para tal, utiliza a infraestrutura da Web Semântica: dados são recursos identificado através de URIs (proporcionando assim meios para identificação única e interligação de dados) e representados através de sentenças expressas na linguagem RDF[39], e de acordo com conceitos especificados em ontologias. Archaeology Data Service Linked Open Data[40] é um exemplo de dados de pesquisa disponibilizados como Dados Abertos e Ligados. Por exemplo, o Governo Federal disponibiliza os dados do orçamento federal[41] em base de dados abertos e ligados, possibilitando a realização de pesquisas científicas que investigam investimentos e custos governamentais.

Serviços de Informação da Pesquisa (SIP) descrevem de forma unificada e integrada os recursos de pesquisa de uma instituição (ou de redes de instituições), como projetos de pesquisa, pesquisadores, laboratórios, instituições, unidades, equipamentos, publicações, resultados/produtos de pesquisas, entre outros, incluindo dados de pesquisa. Esses serviços auxiliam pesquisadores na geração do conhecimento, e gestores na tomada de decisão. Esses serviços representam redes de comunidades de pesquisa em ambientes computacionais, que, com auxílio de ferramentas de visualização, permitem que pesquisadores descubram conteúdos e localizem especialistas e colaboradores potenciais em uma natureza interdisciplinar.

VIVO e DSpace-Cris são plataformas para SIPs. VIVO[42] é uma solução para SIP desenvolvida para a plataforma da Web Semântica. VIVO é usado em várias instituições, como as universidades da Florida[43], Cornell[44] e Carlos III de Madri[45]. DSpace-CRIS[46] é uma extensão do *software* para repositórios digitais DSpace, que relaciona publicações (resultados de pesquisa) com outros tipos de recursos da pesquisa, como projetos, pesquisadores, unidades. É usado, por exemplo, pela Universidade de Tecnologia de Chipre[47].

Em SIPs, dois principais modelos de dados são usados para representar e integrar recursos da pesquisa: VIVO-ISF[48] e CERIF[49]. VIVO-ISF é uma ontologia desenvolvida para representar os recursos da pesquisa no ambiente VIVO, e CERIF é um modelo para representar recursos de pesquisa recomendado pela Comunidade Europeia, que pode

39 Resource Description Framework. https://www.w3.org/RDF/
40 Repositório Archaeology Data Servce - http://data.archaeologydataservice.ac.uk/page/
41 SIOP - Dados Abertos http://orcamento.dados.gov.br/siopdoc/doku.php/acesso_publico:dados_abertos/
42 Goth, G. The Science of Better Science, 2012. DOI: 10.1145/2076450.2076455
43 Vivo na Universidade da Florida -https://vivo.ufl.edu/
44 Vivo na Universidade Cornell - https://scholars.cornell.edu/
45 Vivo na UC3 https://researchportal.uc3m.es/
46 DSpace-CRIS@HKU: Achieving visibility with a CERIF compliant open source system
47 Palmer et. al. DSpace Cris na Cyprus University of Technology, 2014 - http://ktisis.cut.ac.cy/
48 VIVO-ISF Ontology - https://wiki.duraspace.org/display/VTDA/VIVO-ISF+Ontology
49 CERIF - Common European Research Information Format - https://www.eurocris.org/cerif/main-features-cerif

ser usado em DSpace-CRIS. Eagle-I[50] é um exemplo de ambiente para Web Semântica/ Dados Ligados que representa e integra tanto dados de pesquisa quanto recursos de informação de pesquisa.

Este estudo investiga os recursos que as soluções tecnológicas analisadas oferecem para disponibilizar dados no ambiente Dados Abertos e Ligados e integrados com Sistemas de Informação de Pesquisa.

Para a realização do estudo, primeiramente foram elaborados os critérios para analisar as soluções tecnológicas. Esses critérios foram criados a partir das considerações acima apresentadas, considerando, em especial:

- O Modelo de Referência OAIS
- Os Princípios FAIR
- Os Princípios de Citação
- Critérios para Construção de Repositórios Confiáveis
- Representação com Ambiente Linked Data/Web Semântica
- Relações com Sistemas de Informação da Pesquisa
- Desenvolvimento e uso de *Software*

Os critérios foram organizados em categorias. Cada categoria representa um aspecto geral a ser analisado no que diz respeito à solução tecnológica. Essas categorias foram definidas com base em conceitos ou componentes identificados no modelo de referência OAIS[51] e na metodologia desenvolvida pela equipe que criou OAIS[52] que objetiva auxiliar a definição da interface entre o produtor e o repositório, no processo de construção de um repositório.

Foram definidas as seguintes categorias:

- **Representação do Ambiente do Repositório**: Recursos da solução tecnológica para representar o ambiente do repositório e seu modelo (políticas) de funcionamento;
- **Representação dos Conjuntos de Dados**: Prover meios para representar um conjunto de dados (pacote);
- **Descrição e Documentação dos Conjuntos de Dados**: Recursos para descrever e documentar os conjuntos de dados;
- **Produção dos Conjuntos de Dados**: Recursos para a submissão dos dados ao repositório
- **Armazenamento a Longo Prazo e Planejamento da Preservação**: Recursos para o armazenamento seguro e a longo prazo da informação;

50 Eagle-I - https://www.eagle-i.net/

51 CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. Reference model for an open archival information system (OAIS). CSDS 650.0-M-2. 2012 https://public.ccsds.org/pubs/650x0m2.pdf

52 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

- **Acesso e Uso dos Conjuntos de Dados**: Recursos para descoberta dos dados, restringir acesso a dados, entregar dados ao consumidor, prover acesso aos dados;
- **Uso, Desenvolvimento e Manutenção do** *Software:* características do *software* quanto ao seu uso, desenvolvimento e manutanção

As categorias também foram relacionadas aos princípios FAIR e aos critérios para certificação de repositório confiável, Core Trust Seal (CTS). Essa relação indica quais princípios FAIR e quais os critérios de Core Trust Seal estão mais diretamente relacionados às análises conduzidas pela categoria. Isso contribui para a análise de como a solução tecnológica pode dar apoio ao atendimento de princípios FAIR e para a obtenção de um repositório confiável. Os critérios são apresentados na seção 2.

Após a definição dos critérios, as soluções tecnológicas foram analisadas (seção 3). O estudo de cada solução tecnológica foi organizado em seções, cada seção representando uma categoria. Cada categoria foi analisada qualitativamente em que os critérios são citados no texto.

Foram selecionadas para a análise, as soluções tecnológicas que atendem às seguintes características de: serem descritas no diretório de repositório de dados de pesquisa Re3data e, considerando as informações descritas em Re3data, serem as soluções mais usadas, serem usadas no Brasil e serem usadas em pelo menos um repositório que adquiriu certificação de ser repositório confiável.

Os principais conceitos usados na definição das categorias e dos critérios são apresentados no Quadro 1.

Quadro 1 - Conceitos

Produtor	Papel desempenhado pelas pessoas ou sistemas clientes que fornecem as informações a serem preservadas, podendo incluir outros sistemas OAIS [22].
Consumidor	Papel desempenhado por pessoas ou sistemas clientes que interagem com os serviços OAIS para encontrar informações preservadas de interesse e para acessar essas informações em detalhes, podendo incluir outros sistemas OAIS [22].
Comunidade Indicada	Um grupo identificado de potenciais consumidores que devem ser capazes de entender um conjunto específico de informações[22].
Pacote de Informação	Um contêiner lógico composto de Informações de Conteúdo opcionais e Informações de Descrição de Preservação opcionais associadas. Contém informação empacotada usada para delimitar e identificar as Informações de Conteúdo e de Descrição do Pacote usadas para facilitar as pesquisas pelas Informações de Conteúdo.
Pacote de Submissão de Informação	SIP - Submission Information Package Pacote de Informação que contém as informações a serem submetidas a um Repositório OAIS
Pacote de Armazenamento de Informação	AIP - Archival Information Package Pacote de Informação que contém as informações que são armazenadas em um Repositório OAIS

Pacote de Disseminação de Informação	DIP - Dissemination Information Package Pacote de Informação que contém as informações a serem disseminadas a um Repositório OAIS
Metadados	Dados estruturados que provêm informação sobre um ou mais aspectos de outros dados. Podem ser classificados de diferentes formas, tais como descritivos, estruturais e administrativos (técnicos, direitos e de preservação digital).
Metadados Descritivos	Fornecem informações sobre o conteúdo intelectual ou artístico de um objeto. Suportam tarefas específicas do usuário, como descoberta e identificação de conteúdo. Padrões: MARC (registros bibliográficos), Dublin Core (Web), DDI (Dados da Pesquisa – Ciências Sociais e Humanidades)
Metadados Técnicos	Termo genérico para informações técnicas sobre arquivos digitais e objetos multi arquivos, conforme definido por três termos para aspectos importantes de informações técnicas: (1) metadados de características de arquivo para informações técnicas sobre o arquivo digital formatado em questão; (2) metadados de fonte para informações técnicas sobre o item de origem; e (3) metadados de processo para informações sobre os processos técnicos usados para converter o item de origem no arquivo digital [53]
Metadados de Preservação Digital	"Termo fortemente associado ao grupo de trabalho Preservação de Metadados para Materiais Digitais (PREMIS). O grupo definiu um conjunto de metadados de preservação central, suportado por um dicionário de dados, e identificou estratégias para codificar, armazenar e gerenciar esses metadados. Muitos elementos de dados que são importantes para a preservação são encontrados em outras categorias, especialmente aquelas classificadas como administrativas".[54] Padrão: PREMIS[92]
Metadados Estruturais	"No uso pela comunidade de bibliotecas digitais, os metadados estruturais descrevem os elementos intelectuais ou físicos de um objeto digital. Para um arquivo que representa uma única página como um documento composto (por exemplo, um arquivo JPEG 2000 jpm), os metadados estruturais podem incluir informações no layout da página. Em um objeto digital com vários arquivos (por exemplo, um livro digitalizado com muitas imagens de página), os metadados estruturais descrevem os componentes do objeto e seus relacionamentos: páginas, capítulos, índices, índice etc. Esses metadados podem suportar ações sofisticadas de pesquisa e recuperação bem como a navegação e apresentação de objetos digitais". [55] Padrão: METS - Metadata Encoding and Transmission Standard[91]
Estudo	Coleção de arquivos de dados resultantes da coleta intencional de dados por meio de solicitação, observação ou coleta de fontes secundárias para um propósito; utilizando instrumentos e metodologias de coleta descritos; e expressos como arquivos de dados com estruturas lógicas relacionadas ao instrumento de coleta de dados[56].
Conjunto de Dados	Arquivos de dados descritos por uma estrutura lógica. Dados podem se armazenados em um ou mais arquivos de dados.
Livro de Códigos **Codebook**	"Livro que descreve o conteúdo, a estrutura e o layout de uma coleta de dados. Um livro de códigos bem documentado deve conter informações destinadas a ser completas e autoexplicativas para cada variável em um arquivo de dados"[57]

53 FADGI. Glossário - www.digitizationguidelines.gov/term.php?term=metadatatechnical

54 FADGI. Glossário - http://www.digitizationguidelines.gov/term.php?term=metadatapreservation

55 FADGI. Glossário - http://www.digitizationguidelines.gov/term.php?term=metadatastructural

56 DDI Modelo V.4. http://opendatafoundation.org/ddi/srg/Papers/DDIModel_v_4.pdf

57 UCPSR - What is a Codebook? https://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/what-is-a-codebook.html

3 Critérios para análise das soluções tecnológicas

Essa seção apresenta os critérios desenvolvidos para analisar as soluções tecnológicas. O Quadro 2 mostra os critérios organizados em categorias. Neste quadro, junto com a definição de cada categoria, são apresentados todos os critérios de CTS e o princípios FAIR, e são destacados (sublinhados) aqueles princípios e critérios que estão mais diretamente relacionados com à categoria.

Quadro 2 - Critérios

Representação do Ambiente e Modelo de Funcionamento do Repositório

Investigam recursos da solução tecnológica para representar o ambiente do repositório e seu modelo (políticas) de funcionamento.

Exemplo: representação de ambiente de repositório composto por organizações produtoras de conjuntos de dados, suas unidades, subunidades, seus grupos de estudos e os estudos destes grupos; em que unidades podem possuir políticas de submissão e acesso próprias (gestão descentralizada); e que grupos têm autonomia para criar estudos.

Relações com Princípios FAIR e Critérios Core Trust Seal

CTS	
R1	Missão
R2	Licenças
R3	Acesso contínuo, a longo prazo
R4	Confidencialidade e ética – em conformidade com normas
R5	Infraestrutura organizacional
R6	Orientações e feedback de especialistas para manter valor dos dados
R7	Integridade (fixidez, completeza, rastreabilidade, versões) e autenticidade (proveniência, auditoria) dos dados na ingestão, armazenamento e acesso
R8	Dados e metadados aceitos mediante critérios que garante compreensão e relevância aos usuários
R9	Procedimentos documentados para armazenamento
R10	Planejamento da preservação
R11	Dados e metadados com qualidade, e informações para avaliações de qualidade
R12	Atividades de acordo com fluxos de trabalhos definidos
R13	Descoberta e identificação de dados (busca, metadados, harvesting, citação, identificação persistente)
R14	Compressão e uso no futuro, mesmo com mudanças tecnológicas
R15	Infraestrura técnica (*software* e hardware apropriados, maximizando disponibilidade)
R16	Segurança, proteção das instalações e seus dados, produtos, serviços e usuários

FAIR	
F1	Meta(dados) com identificador persistente
F2	Metadados ricos
F3	(Meta)dados usados por recursos de busca
F4	Metadados com identificador do conjunto de dados
A1	Meta(dados) recuperável por identificador
A1.1	Por protocolo de recuperação aberto
A1.2	Autenticação e autorização
A2	Metadados persistem quando dados são removidos
I1	Representação do conhecimento, vocabulários, ontologias
I2	Vocabulários FAIR
I3	Referências a outros meta(dados), contexto
R1	(Meta)dados descritos ricamente por atributos relevantes e precisos
R1.1	Licença
R1.2	Proveniência
R1.2	Padrões relevantes para a comunidade

AMB1	Representação do ambiente no qual os conjuntos de dados estão contidos (ambiente organizado em coleções, estudos, grupos, unidades, subunidades, etc.)
AMB2	Recurso para operacionalizar políticas de funcionamento do ambiente (responsabilidades, grupos de usuários, autenticação de usuários, papéis, autorizações para gerenciar e para realizar funções operacionais do repositório, como submissão e acesso)
AMB3	Recursos para estabelecer políticas descentralizadas, em que unidades, grupos ou estudos têm autonomia de gestão e operação
AMB4	Representação de ambiente integrada com Sistemas de Informação de Pesquisa
AMB5	Representação de ambiente para Web Semântica / Dados Abertos e Ligados

AMB6	Recursos que permitam transparência e *feedback* aos envolvidos, com operações sendo realizadas através de fluxos de trabalhos definidos, permitindo rastreabilidade e auditoria, e com mecanismos de comunicação para manter os envolvidos atualizados. Relatórios e estatísticas.

Representação dos Conjuntos de Dados

Prover meios para representar um conjunto de dados (pacote), considerando suas versões, suas diversas representações (em diversos formatos); cujas estruturas de representação possam ser verificáveis por máquinas a partir de especificações planejadas.

Relações com Princípios FAIR e Critérios Core Trust Seal

CTS		FAIR	
R1	Missão	F1	Meta(dados) com identificador persistente
R2	Licenças	F2	Metadados ricos
R3	Acesso contínuo, a longo prazo	F3	(Meta)dados usados por recursos de busca
R4	Confidencialidade e ética – em conformidade com normas	F4	Metadados com identificador do conjunto de dados
R5	Infraestrutura organizacional	A1	Meta(dados) recuperável por identificador
R6	Orientações e feedback de especialistas para manter valor dos dados	A1.1	Por protocolo de recuperação aberto
R7	**Integridade (fixidez, completeza, rastreabilidade, versões) e autenticidade (proveniência, auditoria) dos dados na ingestão, armazenamento e acesso**	A1.2	Autenticação e autorização
R8	**Dados e metadados aceitos mediante critérios que garante compreensão e relevância aos usuários**	A2	Metadados persistem quando dados são removidos
R9	Procedimentos documentados para armazenamento	I1	Representação do conhecimento, vocabulários, ontologias
R10	Planejamento da preservação	I2	Vocabulários FAIR
R11	**Dados e metadados com qualidade, e informações para avaliações de qualidade**	**I3**	**Referências a outros meta(dados), contexto**
R12	Atividades de acordo com fluxos de trabalhos definidos	**R1**	**(Meta)dados descritos ricamente por atributos relevantes e precisos**
R13	Descoberta e identificação de dados (busca, metadados, harvesting, citação, identificação persistente)	R1.1	Licença
R14	Compressão e uso no futuro, mesmo com mudanças tecnológicas	R1.2	Proveniência
R15	Infraestrutura técnica (*software* e hardware apropriados, maximizando disponibilidade)	**R1.2**	**Padrões relevantes para a comunidade**
R16	Segurança, proteção das instalações e seus dados, produtos, serviços e usuários		

PAC1	Natureza dos conjuntos de dados (Texto, Multimídia, Modelo/Animações, Simulação, *Software*, Específico de Disciplina, Específico de Instrumento, etc.)
PAC2	Estruturas para representar os conjuntos de dados em pacotes, compatíveis com estruturas para representar conjuntos de dados de pesquisa
PAC3	Formatos de arquivos aceitos (gerenciados pela solução tecnológica)
PAC4	Recursos para versionamento de conjuntos de dados
PAC5	Uso de padrões (para pacotes, metadados, formatos de arquivo)
PAC6	Qualidade dos dados. Dados Cinco Estrelas[58]: disponível na web com uma licença aberta (1 estrela), e ainda estruturados e legíveis por máquina (2 estrelas), e ainda em formatos não-proprietários (3 estrelas), e ainda usando padrões da web semântica/ linked data - RDF e URI (4 estrelas), e ainda vinculados com dados de outros, para fornecer o contexto (5 estrelas)

58 5 ★ dos Dados Abertos - https://5stardata.info/pt-BR/

Descrição e Documentação dos Conjuntos de Dados

Recursos para descrever e documentar os conjuntos de dados.

Prover meios para produzir, representar e gerenciar metadados ricos, precisos, indexáveis, úteis ao contexto, aceitos pela comunidade e compreensíveis por máquinas.

Prover **metadados** que permitam a descoberta e uso dos conjuntos de dados (**metadados descritivos**), a obtenção de informações de proveniência e de direitos de uso (**metadados administrativos**), sobre características técnicas dos objetos digitais (**metadados técnicos**), sobre as ações de curadoria digital que foram realizadas (**metadados administrativos**), e descrição dos elementos que compõem o objeto digital e suas estruturas (**metadados estruturais**).

Prover meios para **documentar** como os dados foram produzidos, via metadados e documentos.

Relações com Princípios FAIR e Critérios Core Trust Seal:

CTS	
R1	Missão
R2	Licenças
R3	Acesso contínuo, a longo prazo
R4	Confidencialidade e ética – em conformidade com normas
R5	Infraestrutura organizacional
R6	Orientações e feedback de especialistas para manter valor dos dados
R7	Integridade (fixidez, completeza, rastreabilidade, versões) e autenticidade (proveniência, auditoria) dos dados na ingestão, armazenamento e acesso
R8	Dados e metadados aceitos mediante critérios que garante compreensão e relevância aos usuários
R9	Procedimentos documentados para armazenamento
R10	Planejamento da preservação
R11	Dados e metadados com qualidade, e informações para avaliações de qualidade
R12	Atividades de acordo com fluxos de trabalhos definidos
R13	Descoberta e identificação de dados (busca, metadados, harvesting, citação, identificação persistente)
R14	Compressão e uso no futuro, mesmo com mudanças tecnológicas
R15	Infraestrura técnica (*software* e hardware apropriados, maximizando disponibilidade)
R16	Segurança, proteção das instalações e seus dados, produtos, serviços e usuários

FAIR	
F1	Meta(dados) com identificador persistente
F2	Metadados ricos
F3	(Meta)dados usados por recursos de busca
F4	Metadados com identificador do conjunto de dados
A1	Meta(dados) recuperável por identificador
A1.1	Por protocolo de recuperação aberto
A1.2	Autenticação e autorização
A2	Metadados persistem quando dados são removidos
I1	Representação do conhecimento, vocabulários, ontologias
I2	Vocabulários FAIR
I3	Referências a outros meta(dados), contexto
R1	(Meta)dados descritos ricamente por atributos relevantes e precisos
R1.1	Licença
R1.2	Proveniência
R1.2	Padrões relevantes para a comunidade

DOC1	Informação de proveniência e de contextualização da produção dos dados, incluindo versionamento, produtores, processos de produção, recursos relacionados (como publicações), compreensível por máquina e humanos e aceita pela comunidade (metadados administrativos)
DOC2	Informação descritiva de ações de gestão e preservação digital, incluindo verificação da integridade do material, compreensível por máquina e humanos e aceita pela comunidade (metadados administrativos)
DOC3	Informação descritiva sobre o conteúdo intelectual, compreensível por máquina e humanos e aceita pela comunidade (metadados descritivos)
DOC4	Informação descritiva sobre a estrutura de representação dos dados – pacotes de informação compreensível por máquina e aceita pela comunidade (metadados estruturais)
DOC5	Informação descritiva sobre aspectos técnicos dos objetos digitais (formatos de arquivos, versões dos formatos), compreensível por máquina e aceita pela comunidade (metadados técnicos)
DOC6	Vocabulários controlados (listas de termos, classificações, tesauros)
DOC7	Recursos para representar novos esquemas de metadados e estender esquemas existentes
DOC8	Recursos para realizar mapeamentos (*crosswalks*) entre esquemas de metadados ou gerenciar metadados representados em múltiplos formatos

DOC9	Descrever os conjuntos de dados em ambiente Linked Data/Web Semântica
DOC10	Descrever os conjuntos de dados integrados com Sistemas de Informação da Pesquisa
DOC11	Documentação dos dados

Produção dos Conjuntos de Dados – Relação com Produtor

Prover meios para submissão dos dados ao repositório, observando pacotes de submissão, fluxo de submissão, funções que verificam a autenticidade do produtor e se o material submetido está íntegro (verificação) e em conformidade (validação) com as especificações planejadas (estrutura do pacote, metadados, proveniência, planejamento da preservação), e geração do pacote de armazenamento a partir do pacote submetido pelo produtor.

Microserviços de preservação digital[59] que atuam no pacote de submissão: verificação (integridade, fixidez, checagem de vírus), normalização (conversão para formatos aceitos), validação (conformidade com especificações, como DROID[60]), caracterização (extração de metadados técnicos, via aplicativos como JHOVE[61]), transcrição (OCR), etc.

Prover meios para gerenciar acordos de submissão e licenças.

Relações com Princípios FAIR e Critérios Core Trust Seal

CTS		FAIR	
R1	Missão	F1	Meta(dados) com identificador persistente
R2	Licenças	F2	Metadados ricos
R3	Acesso contínuo, a longo prazo	F3	(Meta)dados usados por recursos de busca
R4	Confidencialidade e ética – em conformidade com normas	F4	Metadados com identificador do conjunto de dados
R5	Infraestrutura organizacional	A1	Meta(dados) recuperável por identificador
R6	Orientações e feedback de especialistas para manter valor dos dados	A1.1	Por protocolo de recuperação aberto
R7	**Integridade (fixidez, completeza, rastreabilidade, versões) e autenticidade (proveniência, auditoria) dos dados na ingestão, armazenamento e acesso**	**A1.2**	**Autenticação e autorização**
R8	Dados e metadados aceitos mediante critérios que garante compreensão e relevância aos usuários	A2	Metadados persistem quando dados são removidos
R9	Procedimentos documentados para armazenamento	I1	Representação do conhecimento, vocabulários, ontologias
R10	Planejamento da preservação	I2	Vocabulários FAIR
R11	Dados e metadados com qualidade, e informações para avaliações de qualidade	I3	Referências a outros meta(dados), contexto
R12	**Atividades de acordo com fluxos de trabalhos definidos**	R1	(Meta)dados descritos ricamente por atributos relevantes e precisos
R13	Descoberta e identificação de dados (busca, metadados, harvesting, citação, identificação persistente)	**R1.1**	**Licença**
R14	Compressão e uso no futuro, mesmo com mudanças tecnológicas	**R1.2**	**Proveniência**
R15	Infraestrura técnica (*software* e hardware apropriados, maximizando disponibilidade)	R1.2	Padrões relevantes para a comunidade
R16	Segurança, proteção das instalações e seus dados, produtos, serviços e usuários		

SUB1	Gerenciamento de Acordos de Submissão e Licença (definição de acordos de submissão e licenças, gerenciamento de acordos e licenças acordadas).
SUB2	Fluxos de submissão: produção, aprovação de pacotes de submissão. Publicação de pacotes submetidos e aceitos
SUB3	Transferência legal da custódia dos dados ao repositório, autenticação do produtor e usuários envolvidos, aprovação e armazenamento de acordos de submissão, certificação e registros de proveniência.
SUB4	Validação e verificação pacotes de submissão (microserviços), verificação e validação de identificadores globais persistentes, geração dos pacotes de armazenamento para submissões aceitas

59 Abrams, S. et al.: Curation Micro-Services: A Pipeline Metaphor for Repositories. 2011 - https://journals.tdl.org/jodi/index.php/jodi/article/view/1605/1766

60 DROID – Ferramenta que Identifica o formato preciso de um arquivo - http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

61 JHOVE - Ferramenta para Identificação, Validação e Caracterização de Formatos http://jhove.openpreservation.org/

SUB5	Extração, verificação e/ou produção de metadados técnicos, administrativos e descritivos. Registro em metadados das ações de submissão.
SUB6	Submissão em lote. Submissão por máquina e em ambientes distribuídos, recepção de pacotes submetidos por outros ambientes e/ou depósito de materiais aprovados em outros ambientes, via protocolos e estruturas de pacotes de submissão aceitos e padronizados, como Sword[62], OAI-PMH[63] (alimentação da BD via colheita, BD é agregador), OAI-OR[E64]
SUB7	Uso de padrões (para pacotes e para protocolos de submissão)

Armazenamento a Longo Prazo - Planejamento da Preservação

Prover meios para armazenamento seguro da informação, em pacotes de armazenamento, em conformidade com as especificações planejadas; para checagem da integridade das informações; com o registro das ações de preservação digital e seus efeitos e o armazenamento de objetos digitais decorrentes dessas ações (como novas representações decorrentes de migrações).

Prover meios para que o planejamento da preservação trabalhe de forma integrada com os serviços oferecidos pelo repositório, como por exemplo, gerenciar formatos, monitorar formatos usados, microserviços de curadoria digital (verificação, replicação, migração, etc.)

Relações com Princípios FAIR e Critérios Core Trust Seal:

CTS		**FAIR**	
R1	Missão	F1	Meta(dados) com identificador persistente
R2	Licenças	F2	Metadados ricos
R3	**Acesso contínuo, a longo prazo**	F3	(Meta)dados usados por recursos de busca
R4	Confidencialidade e ética – em conformidade com normas	F4	Metadados com identificador do conjunto de dados
R5	Infraestrutura organizacional	A1	Meta(dados) recuperável por identificador
R6	Orientações e feedback de especialistas para manter valor dos dados	A1.1	Por protocolo de recuperação aberto
R7	**Integridade (fixidez, completeza, rastreabilidade, versões) e autenticidade (proveniência, auditoria) dos dados na ingestão, armazenamento e acesso**	A1.2	Autenticação e autorização
R8	Dados e metadados aceitos mediante critérios que garante compreensão e relevância aos usuários	**A2**	**Metadados persistem quando dados são removidos**
R9	**Procedimentos documentados para armazenamento**	I1	Representação do conhecimento, vocabulários, ontologias
R10	**Planejamento da preservação**	I2	Vocabulários FAIR
R11	Dados e metadados com qualidade, e informações para avaliações de qualidade	I3	Referências a outros meta(dados), contexto
R12	Atividades de acordo com fluxos de trabalhos definidos	R1	(Meta)dados descritos ricamente por atributos relevantes e precisos
R13	Descoberta e identificação de dados (busca, metadados, harvesting, citação, identificação persistente)	R1.1	Licença
R14	**Compressão e uso no futuro, mesmo com mudanças tecnológicas**	R1.2	Proveniência
R15	Infraestrura técnica (*software* e hardware apropriados, maximizando disponibilidade)	R1.2	Padrões relevantes para a comunidade
R16	Segurança, proteção das instalações e seus dados, produtos, serviços e usuários		

PD1	Serviços/microserviços para garantir acesso a longo prazo: Gerenciamento de armazenamento, checagem de integridade e ações de preservação digital, como refrescamento (troca de mídia), replicação (cópias de segurança), reempacotamento (reestruturações de pacotes), transformação (migração).
PD2	Planejamento e ações de preservação digital: gerenciamento de formatos de arquivos (políticas para formatos aceitos), planejamento e execução automatizada de ações planejadas (especificação de planos e execução de microserviços quando objetos submetidos ou armazenados não estão em conformidade com os planos atuais), informações de apoio à preservação digital (estatísticas de uso dos formatos, mídias, etc),

62 Protocolo Sword. http://swordapp.org/about/

63 Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh/

64 Open Archives Initiative Object Exchange and Reuse. https://www.openarchives.org/ore/

PD3	Integração com serviços de preservação digital de terceiros ou colaborativos, como Lockss[65], armazenamento nas nuvens, etc. Integração com serviços ou infraestruturas locais de preservação digital, como iRods[66] e Archivematica[67]
PD4	Exclusão de dados, com manutenção dos metadados, autorizações e registro da ação
PD5	Usado em Repositório Digital Confiável certificado

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, como proveniência e licenças, através de protocolos de comunicação padronizados, abertos e gratuitos.

Prover meios para identificação única, persistente e global para os conjuntos de dados, considerando versionamento de dados e conjuntos de dados sendo disponibilizados em várias representações.

Prover meios para identificação única e persistente dos atores envolvidos na proveniência dos dados, como produtores de dados.

Prover meios para permitir a citação dos dados, via metadados de citação e identificadores persistentes que levam a esses metadados e a outros documentos, atribuindo créditos aos produtores, disponibilizando informações de proveniência e que permitem a verificação de fixidez, considerando que os dados em questão podem ser versões, séries, subconjuntos.

Relações com Princípios FAIR e Critérios Core Trust Seal:

CTS	**FAIR**
R1 Missão R2 Licenças **R3 Acesso contínuo, a longo prazo** R4 Confidencialidade e ética – em conformidade com normas R5 Infraestrutura organizacional R6 Orientações e feedback de especialistas para manter valor dos dados R7 Integridade (fixidez, completeza, rastreabilidade, versões) e autenticidade (proveniência, auditoria) dos dados na ingestão, armazenamento e acesso R8 Dados e metadados aceitos mediante critérios que garante compreensão e relevância aos usuários R9 Procedimentos documentados para armazenamento R10 Planejamento da preservação R11 Dados e metadados com qualidade, e informações para avaliações de qualidade R12 Atividades de acordo com fluxos de trabalhos definidos **R13 Descoberta e identificação de dados (busca, metadados, harvesting, citação, identificação persistente)** R14 Compressão e uso no futuro, mesmo com mudanças tecnológicas R15 Infraestrura técnica (*software* e hardware apropriados, maximizando disponibilidade) R16 Segurança, proteção das instalações e seus dados, produtos, serviços e	**F1 Meta(dados) com identificador persistente** F2 Metadados ricos **F3 (Meta)dados usados por recursos de busca** **F4 Metadados com identificador do conjunto de dados** A1 Meta(dados) recuperável por identificador A1.1 Por protocolo de recuperação aberto A1.2 Autenticação e autorização A2 Metadados persistem quando dados são removidos I1 Representação do conhecimento, vocabulários, ontologias I2 Vocabulários FAIR I3 Referências a outros meta(dados), contexto R1 (Meta)dados descritos ricamente por atributos relevantes e precisos R1.1 Licença R1.2 Proveniência R1.2 Padrões relevantes para a comunidade

AC1	Recuperação de informação (busca por metadados, busca por ocorrência de palavras, navegação por facetas (assunto, título, produtores, tipos de dados, unidades, subunidades, grupos, estudos, etc)
AC2	Informações sobre direitos de uso, licenças
AC3	Informações de proveniência e para uso dos dados (metadados descritivos, metadados de proveniência, documentação)
AC4	Informações de citação

65 LOCKSS - Lots Of Copies Keep Stuff Safe - https://www.lockss.org/

66 *Software* iRODS - https://irods.org/

67 *Software* Archivematica - https://www.archivematica.org

AC5	Identificadores globais e persistentes para conjuntos de dados, considerando versionamento e objetos digitais em várias representações, e usando serviços, protocolos e padrões aceitos pela comunidade (como os padrões DOI [30] e Handle System [31] e os serviços DataCite [34] etc)
AC6	Identificadores globais e persistentes para recursos relacionados aos conjuntos de dados e usando serviços, protocolos e padrões aceitos pela comunidade (como os ORCID[68], para identificar unicamente pessoas/pesquisadores)
AC7	Acesso aos dados e aos metadados via identificador e protocolo aberto
AC8	Restrições de acesso (nível de arquivo, conjunto de dados, grupos, unidades, etc., a determinados grupos), embargos, acesso mediante registro do usuário, registro de solicitações de uso e relacionamento com usuários (Livro de Visitas / *Guestbook*)
AC9	Gerenciamento e autenticação de usuários envolvidos com acesso
AC10	Entrega de dados ao consumidor (pessoa ou sistema) nas representações (formatos e estruturas) adequadas, acesso via API
AC11	Entrega dos metadados ao consumidor (pessoa ou sistema) nas representações (formatos e estruturas) adequadas, protocolos de colheita de metadados (OAI-PMH [63], Z39-50[69]), acesso via API
AC12	Ferramentas para visualização e análise de dados
AC13	Estatísticas e relatórios de uso
AC14	Acesso às descrições em formatos para Linked Data/Web Semântica
AC15	Recuperação em ambiente Linked Data/Web Semântica (endpoints, SPARQL)

Desenvolvimento, Manutenção e Uso do *Software*	
SW1	Tecnologia e Plataforma
SW2	Distribuição e Versionamento
SW3	Estratégia de Desenvolvimento do *Software*
SW4	Licença de Uso
SW5	Desempenho e Escalabilidade
SW6	Presença – Usuários, Uso no Brasil

68 ORCID - https://orcid.org/

69 Protocolo Z 39.50 - https://www.loc.gov/z3950/agency/

4 Soluções Tecnológicas

O Quadro 3 apresenta as principais soluções tecnológicas usadas para repositórios de dados de pesquisa, cadastradas no diretório de repositórios de dados Re3Data[70]. A análise das soluções tecnológicas investiga em profundidade DSpace (seção 4.1) e Dataverse (seção 4.2), por estas serem as soluções mais usadas, serem usadas no Brasil e por serem adotadas por repositórios que obtiveram certificação de repositório confiável. DSpace e Dataverse são analisadas com relação às características e funcionalidades apresentadas pela distribuição oficial do *software*, excluindo extensões desenvolvidas como experimentos ou para solucionar requisitos específicos de um repositório.

Quadro 3 - Uso de Soluções Tecnológicas em Repositórios de Dados

Solução Tecnológica	Uso em Repositório de Dados	Repositórios Confiáveis Certificados por Trust Core Seal ou Data Seal Approval
Dataverse	69	3
DSpace	62	11
CKAN	52	0
Fedora	31	14
Eprints	31	0

Fonte: Diretório Re3Data, em 15/11/2018

4.1 DSpace

Esta seção detalha o estudo da solução tecnológica DSpace.

4.1.1 Desenvolvimento, Manutenção e Uso do *Software*

Essa seção analisa DSpace com relação aos critérios: tecnologia e plataforma [SW1], distribuição e versionamento [SW2], estratégia de desenvolvimento do *software* [SW3], licença de uso [SW4], desempenho e escalabilidade [SW5], Presença – Usuários, Uso no Brasil [SW6].

DSpace foi lançado como um esforço conjunto entre os desenvolvedores do Massachusetts Institute of Technology (MIT) Libraries e a Hewlett-Packard (HP) Labs. Com o crescimento da comunidade de usuários, HP e MIT formaram em conjunto a DSpace Foundation, uma organização sem fins lucrativos que forneceu liderança e suporte. 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 [SW1][71]. Entre os principais requisitos para o funcionamento do *software* destacam-se:

70 Diretório de Repositórios de Dados Re3data - https://www.re3data.org/browse

71 DSpace - https://duraspace.org/dspace/download/

- Java JDK 7 ou 8 (OpenJDK[72] ou Oracle JDK[73]);
- Apache Maven[74] 3.0.5 ou posterior (3.3.9+ para compilar o tema Mirage 2);
- Apache Ant[75] 1.8 ou superior;
- PostgreSQL[76] 9.4+ (with pgcrypto) ou Oracle[77] 10g+;
- Apache Tomcat[78] 7.0.30+ ou 8.0.33+, ou Servlet Engine equivalente;
- Git[79]

A configuração de *hardware* recomendada para um sistema em produção inclui pelo menos 8GB de memória RAM, principalmente devido ao consumo do sistema gerenciador de banco de dados relacional e do servidor Java web. São recomendados pelo menos 200 GB de espaço de armazenamento, distribuídos em discos de alta *performance,* configurados em RAID.

DSpace adota o seguinte esquema de distribuição versionamento do *software* [SW2]. Versões principais (*major releases*) podem incluir novas funcionalidades, melhorias de sistema, mudanças arquiteturais e correção de falhas. Versões menores (*minor releases*) são geradas a partir de correção de falhas (*bug fixes*) em versões principais. O esquema de numeração das versões consiste em [*major*].[*minor*]. Em 2018, a versão atual disponível é a 6.3. 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.

DSpace é distribuído sob licença BSD (Berkeley *Software* Distribution), portanto pode ser redistribuído com ou sem modificação, mas usa muitas bibliotecas de terceiros regidas por diferentes tipos de licença[80], que devem ser observadas [SW4].

Características de desempenho e escalabilidade do DSpace são bastante dependentes da configuração do servidor Java web, do sistema gerenciador de banco de dados relacional e do sistema de recuperação de informações [SW5]. É essencial permitir que Tomcat possa consumir uma quantidade de memória maior que o definida por padrão. PostgreSQL ou Oracle Database devem ser sintonizados para admitirem *buffers* compartilhados de tamanho adequado, *checkpoints* de escrita dos *logs* em momentos oportunos, entre outras várias configurações.

Por fim, existem 117 instalações de DSpace no Brasil registradas na comunidade Duraspace[81] [SW6], principalmente em instituições acadêmicas e governamentais tais como Universidades, Bibliotecas Digitais e Tribunais de Justiça. Entretanto, a grande maioria destas instalações não são dedicadas a repositórios de dados, mas de documentos. Na comunidade Re3data [68] estão registradas apenas duas instalações de repositórios de dados brasileiras.

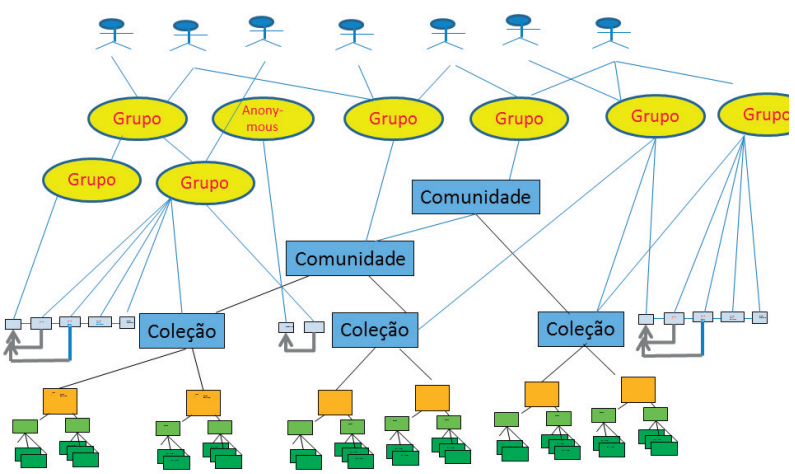
72 OpenJDK - http://openjdk.java.net/
73 Oracle JDK- https://www.oracle.com/technetwork/java/javase/downloads/index.html
74 Apache Maven - http://maven.apache.org/download.html
75 Apache Ant - http://ant.apache.org/
76 PostgreSQL - http://www.postgresql.org/
77 Oracle Database - http://www.oracle.com/database/
78 Apache Tomcat - http://tomcat.apache.org/whichversion.html
79 Git - https://git-scm.com/downloads
80 Licença de DSpace- https://duraspace.org/dspace/license/
81 Comunidade Duraspace - https://duraspace.org/registry/register-your-site/

4.1.2 Representação do Ambiente do Repositório

Essa seção analisa DSpace com relação aos critérios: representação do ambiente [AMB1], recursos para operacionalizar políticas de funcionamento do ambiente [AMB2], recursos para estabelecer políticas descentralizadas e autônomas [AMB3], representação de ambiente integrada com Sistemas de Informação de Pesquisa [amb4], representação de ambiente para Web Semântica / Dados Abertos e Ligados [AMB5], recursos que permitam transparência e feedback aos envolvidos [AMB6].

DSpace possui recursos que permitem variadas configurações para ambientes de repositório. Comunidades são os recursos que DSpace oferece para representar entidades organizacionais, como organizações, unidades, subunidades e grupos [AMB1] (Figura 1). Comunidades podem conter subcomunidades e coleções. Coleções representam conjuntos de itens.

Figura 1 - Comunidades e Coleções

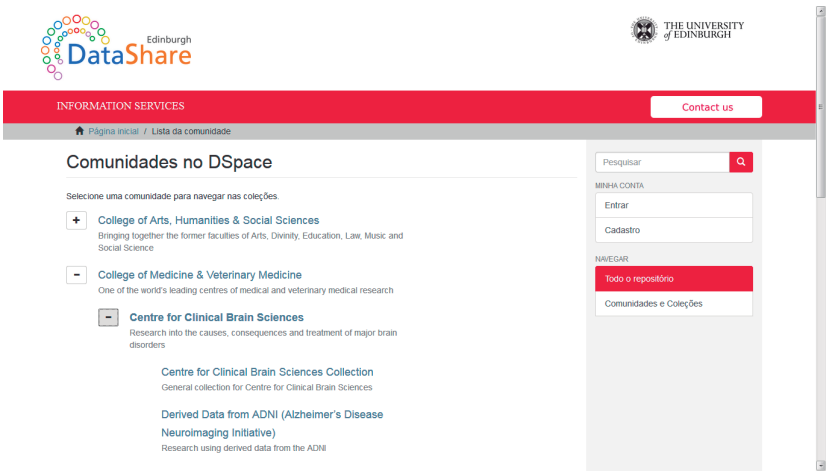


Fonte: Dados da pesquisa

DSpace não foi desenvolvido para dados de pesquisa, entretanto, comunidades podem ser usadas para representar instituições produtoras ou custodiadoras de conjuntos de dados, e/ou grupos de pesquisa. Por exemplo, DataShare [3], repositório de dados da Universidade de Edinburg, usa comunidades e coleções para representar hierarquicamente a comunidade de pesquisa da Universidade[82] [AMB1] (Figura 2).

82 Comunidades de Datashare - https://datashare.is.ed.ac.uk/community-list

Figura 2 - DataShare - Comunidades e Coleções



Fonte: DataShare - https://datashare.is.ed.ac.uk/

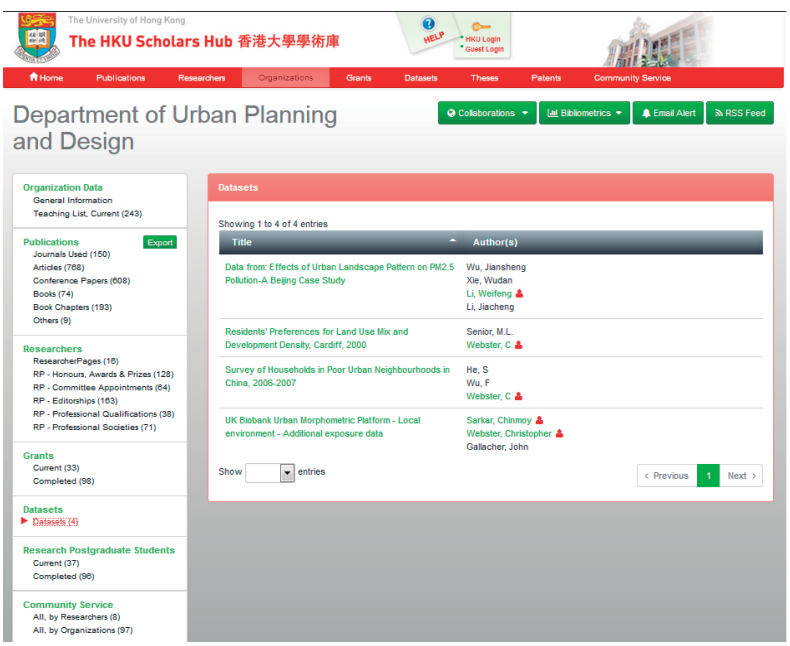
DSpace dispões de recursos para implementar comunidades e coleções com políticas de funcionamento próprias e distintas [AMB2] [AMB3]. Permite o registro e autenticação de usuários; a organização de usuários em grupos; e a atribuição de autorizações para grupos e usuários [AMB2] (Figura 1). Possibilita a definição de políticas em que somente grupos autorizados podem acessar comunidades, coleções e itens, assim como submeter item e administrar comunidades, coleções, fluxos de submissões e itens (alterar, criar, remover) [AMB2]. Isso possibilita a criação de políticas descentralizadas, em que comunidades podem ter autonomia para criar e definir políticas de acesso e submissão, assim como criação e gerenciamento de subcomunidades e coleções [AMB3]. Somente usuários do grupo administrador podem gerenciar usuários, esquemas de metadados e políticas de formatos.

DSpace não foi desenvolvido para ser um Sistema de Informação de Pesquisa [AMB4]. Em DSpace não é possível a representação de recursos de Sistemas de Informação de Pesquisa, como projetos, pessoas/pesquisadores, unidades organizacionais, equipamentos, laboratórios, etc. Unidades organizacionais podem ser representadas em DSpace através de comunidades, entretanto, estas são somente vistas pelo *software* como gestoras e possuidoras de conjuntos de dados, não dispondo de propriedades para descrição de unidades organizacionais, nem propriedades que as relacionam com outros recursos de um Sistema de Informação de Pesquisa.

DSpace-CRIS[40] [AMB4] é uma extensão de DSpace para atender necessidades de um Sistema de Informação da Pesquisa. DSpace-CRIS permite a representação e a interligação de pesquisadores, unidades organizacionais, projetos (*grants*) e resultados de pesquisa (publicações, patentes, teses). Possibilita a definição de outras classes arbitrárias de objetos, incluindo a especificação de propriedades e de relações com outros tipos de objetos, assim como a definição de formulários para descrição desses objetos. Foi desenvolvido com base no padrão Europeu para representação de Sistemas de Informação da Pesquisa, CERIF[43], permitindo a construção de sistemas compatíveis com CERIF.

DSpace-CRIS é usado no repositório da Universidade de Hong Kong[83]. Nesse repositório, conjuntos de dados são representados através da classe *Dataset*. Essa classe contém propriedades que descrevem esses conjuntos e que estabelecem relações com outros objetos do SIP e arquivos. A Figura 3 mostra a entidade organizacional **Department of Urban Planning and Design**, com ligações a outros recursos, como publicações (representados como itens de DSpace), pesquisadores, projetos/*grants* e conjuntos de dados.

Figura 3 - HKU - Unidade Organizacional e Entidades Relacionadas



Fonte: Repositório HKU - http://hub.hku.hk/cris/ou/ou00145

DSpace possui (a partir da versão 5) um módulo de integração com a Web Semântica [AMB5]. Através deste módulo, metadados de um conteúdo armazenado em DSpace são convertidos e armazenados em uma BD de triplas RDF (triple store/endpoint) imediatamente após o conteúdo ser criado ou atualizado[84], permitindo a busca e a interoperabilidade semântica via SPARQL. DSpace permite ainda que metadados sejam colhidos via protocolo OAI-PMH no formato da Web Semântica, isto é, na representação RDF/XML. Para tal, os metadados representados em Dublin Core são convertidos para RDF/XML.

Com relação à transparência e *feedback* [AMB6], DSpace possui recursos baseados em fluxo de trabalho (*workflow*) para submissão de materiais (Figura 9, seção 4.1.5), no qual produtores constroem pacotes de submissão de informação, e usuários

83 Repositório HKU - http://hub.hku.hk

84 DSpace 6.x Documentation. DSpace Linked (Open) Data - https://wiki.duraspace.org/display/DSDOC6x/Linked+%28Open%29+Data

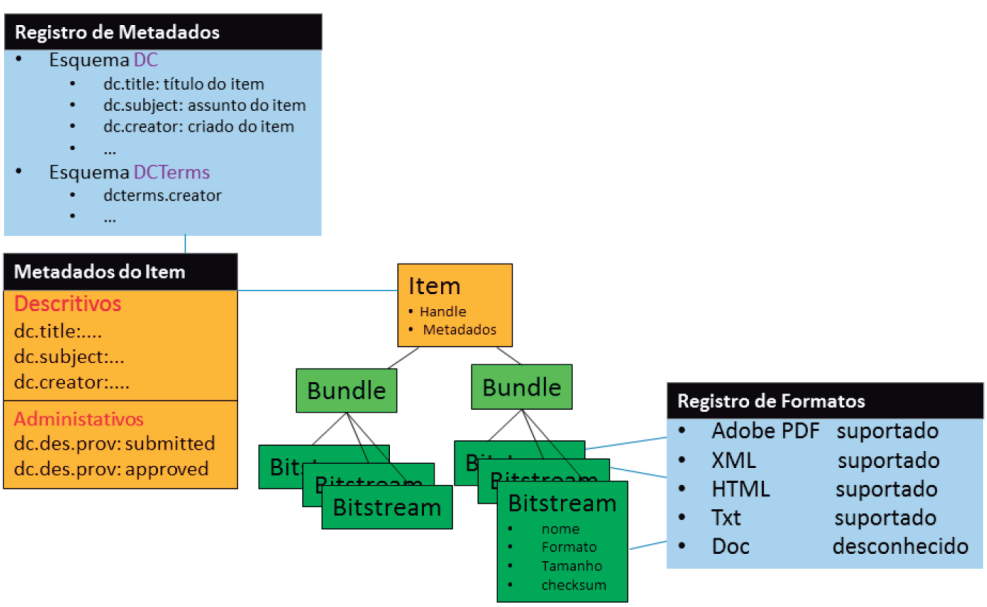
podem ser devidamente configurados para aprovar submissões e corrigir os metadados dos pacotes submetidos. O produtor e os demais envolvidos no fluxo de submissão têm acesso ao andamento do processo. DSpace registra, na forma de metadados de proveniência, os resultados de cada passo do fluxo de submissão (submissão, avaliação, verificação), assim como o evento em que o pacote foi publicado. DSpace também possui relatório de estatísticas que mostra o uso do item.

4.1.3 Representação dos Conjuntos de Dados

Essa seção analisa DSpace com relação aosd critérios: natureza dos conjuntos de dados [PAC1], estruturas para representar os conjuntos de dados [PAC2], formatos de arquivos [PAC3], recursos para versionamento de conjuntos de dados [PAC4], dados cinco estrelas [PAC5], uso de padrões da web semântica/linked data [PAC6].

Para representar conjuntos de dados, DSpace dispõe de itens que são armazenados em coleções. Uma coleção representa um conjunto de itens com mesmas características. Todos os itens de uma coleção são descritos por um mesmo conjunto de metadados, passam por um mesmo fluxo de submissão e estão sujeitos às mesmas políticas de submissão, gestão e acesso. Um item em DSpace é uma estrutura complexa, composta por *bundles*, que são espécies de pastas, e *bitstreams*, que são arquivos binários [PAC2] (Figura 4). Tendo como base a terminologia OAIS, pode-se dizer que DSpace usa *bundle* e *bitstreams*, e seus metadados, para representar Pacotes de Armazenamento de Informação (AIPs)

Figura 4 – Estrutura de um Item em DSpace



Fonte: Dados da pesquisa

Por meio de um item é possível a representação de um objeto complexo composto por arquivos (*bitstreams*) que podem ser organizados em pastas (*bundles*). En-

tretanto, pastas não podem conter outras pastas. DSpace permite a representação de arquivos (*bitstreams*) nos mais variados formatos, incluindo texto (como txt, pdf, csv), imagem (como jpg e tiff), vídeo (como mpeg) e som (wave). Os formatos usados devem ser cadastrados no módulo chamado Registro de Formatos (Figura 6) [PAC1] [PAC3]. Dessa forma, DSpace controla e permite o gerenciamento de formatos.

Em DSpace, grupo de pesquisa/unidade, estudo e conjunto de dados podem ser representados da seguinte forma:

- **Unidade**: Comunidade de DSpace
- **Grupo de pesquisa**: Comunidade ou Coleção de DSpace
- **Estudo**: Item de DSpace
- **Conjunto de Dados de estudo**: arquivos (*biststreams*) pertencentes a um item, organizados em pastas (*bundles*).

Por exemplo, a Figura 5 apresenta um estudo no repositório Datashare. Este estudo (**UK Statutury Instuments**) é representado como item.

Figura 5 - Estudo em DataShare

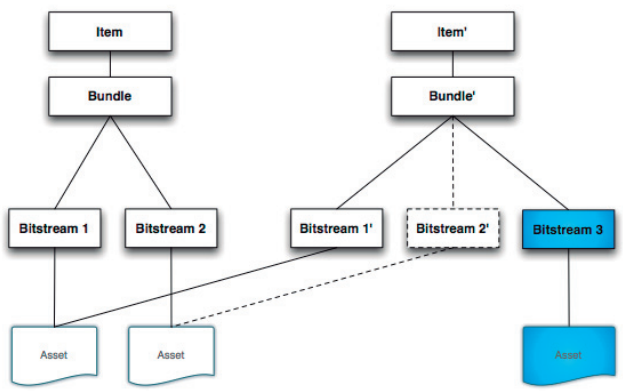
Fonte: Datashare - https://datashare.is.ed.ac.uk/handle/10283/774

Para submissão de item via interface Web, o limite de cada arquivo (*bitstream*) é 2GB, devido a restrições da linguagem utilizada para o desenvolvimento da interface. Globus é um exemplo de serviço de Big Data que usa DSpace. Nesse caso, DSpace é usado somente como serviço de publicação[85]. Um módulo especial de submissão foi desenvolvido para esse repositório.

85 Kyle Chard. Globus Data Publication as a Service: Lowering Barriers to Reproducible Science. 2015 https://www.globus.org/sites/default/files/Data_Publication_Lowering_Barriers.pdf

Um módulo de DSpace [86] foi criado para prover serviço de versionamento de item [PAC 4], sendo incorporado na distribuição de DSpace. Para cada nova versão de um item (como no caso da inclusão de um novo *bitstream*), 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 (Figura 6).

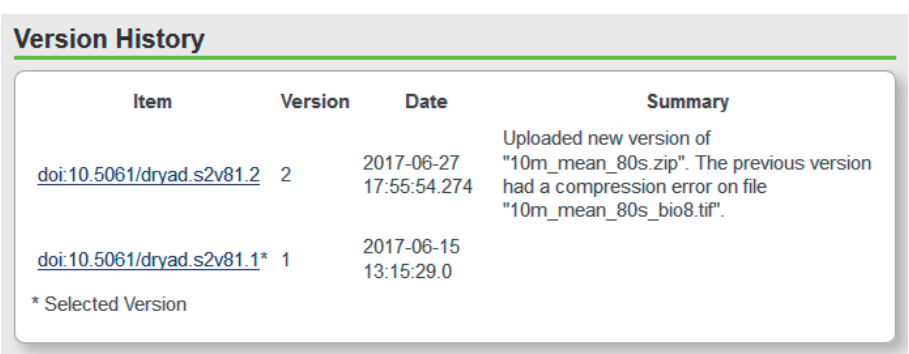
Figura 6 – Versionamento de Item em DSpace



Fonte: DSpace 6.x Documentation - https://wiki.duraspace.org/display/DSDOC6x/Item+Level+Versioning

Por exemplo, no Repositório Dryad[87], a atualização de algum arquivo de um item leva a uma nova versão, e cada arquivo dessa versão possui um novo identificador (Figura 7).

Figura 7 – Versionamento em Dryad



Fonte: Dryad - https://datadryad.org/resource/doi:10.5061/dryad.s2v81.2

Um fator importante para a preservação digital é representar itens em pacotes que são independentes do *software* e que ficam armazenados em uma mesma localização/estrutura no disco. Dessa forma, a obtenção e a preservação de um item não fica dependente a preservação do *software* que constrói e gerencia este item. DSpace representa itens de forma dependente do *software*. Arquivos (*bitstreams*) são armazenados no

86 DSpace 6.x Documentation .Using DSpace https://wiki.duraspace.org/display/DSDOC6x/Item+Level+Versioning

87 Repositório Dryad - https://datadryad.org/

sistema de arquivos, e a estrutura do item (seus metadados e *bundles*) são representados em banco de dados relacional. Essa estratégia torna a obtenção do item dependente do *software*, trazendo dificuldades para desenvolver estratégias de preservação digital. Entretanto, para fins de preservação digital, DSpace permite a importação/exportação de pacotes em estruturas que seguem padrões e recomendações de preservação digital. Pacotes são representados através de estruturas de pastas, com o uso dos padrões METS para descrever as estruturas dos objetos (metadados estruturais)[88] [PAC5].

DSpace permite a publicação de dados em formatos não proprietários e legíveis por máquina (dados 3 estrelas). Também possibilita o armazenamento de arquivos no formato RDF/XML, isto é, arquivos que contém dados representados no padrão da Web Semântica (4 estrelas) e interligados (5 estrelas). Não permite o armazenamento, o gerenciamento e a manipulação de triplas RDF, pois não armazena e gerencia triplas RDF (triple store) [PAC6]. Metadados podem ser armazenados em bases de dados de triplas.

4.1.4 Descrição e Documentação dos Conjuntos de Dados

Essa seção analisa DSpace com relação a: informação de proveniência e de contextualização da produção dos dados [DOC1], informação descritiva de ações de gestão e de preservação digital (metadados administrativos) [DOC2], informação descritiva sobre o conteúdo intelectual (metadados descritivos) [DOC3], informação descritiva sobre a estrutura de representação dos dados (metadados estruturais) [DOC4], informação descritiva sobre aspectos técnicos dos objetos digitais (metadados técnicos) [DOC5], vocabulários controlados [DOC6], representar novos esquemas de metadados e estender esquemas existentes [DOC7], realizar mapeamentos (crosswalks) entre esquemas de metadados ou gerenciar metadados representados em múltiplos formatos [DOC8], descrever os conjuntos de dados em ambiente Linked Data/Web Semântica [DOC9], descrever os conjuntos de dados integrados com Sistemas de Informação da Pesquisa [DOC10], documentação dos dados [DOC11].

DSpace permite a definição de esquemas de metadados, com elementos não hierárquicos (elementos não podem conter outros elementos). Representa informações de formato e proveniência/preservação digital através de metadados, e representa metadados estruturais do pacote e outras informações na base de dados relacional do ambiente.

Para descrever coleções e comunidades, DSpace disponibiliza um conjunto fixo e limitado de metadados, que são armazenados na base de dados relacional que dá apoio ao funcionamento do repositório. Para descrever itens, DSpace permite a criação de novos elementos, e a organização desses elementos em esquemas (em Registro de Formatos - Figura 4) [DOC7]. Dessa forma, DSpace possibilita a definição de esquemas de metadados que atendam às necessidades da comunidade do repositório.

88 DSpace 6.x Documentation - System AdministratiDSpace AIP Format - https://wiki.duraspace.org/display/DSDOC6x/DSpace+AIP+Format

Em DSpace, metadados são representados na forma propriedade-valor. Por isso, o *software* não permite a representação de elementos com estruturas complexas, isto é, a representação de elementos cujos conteúdos são também desdobrados em novos elementos. Isso traz dificuldades na configuração da ferramenta para representar esquemas de metadados que são especificados em XML e que contém elementos em estrutura hierárquica, como DDI[48] e DataCite[62]. Entretanto, DSpace dispõe de mecanismos de mapeamento de metadados, via regras XSLT, que permite a geração de representações de metadados em estruturas hierárquicas XML a partir de representações não hierárquicas. Por exemplo, metadados na forma propriedade-valor, como nome e afiliação de uma pessoa, podem ser mapeados para uma estrutura hierárquica que representa uma pessoa com suas propriedades.

Com relação a metadados descritivos [DOC3], DSpace tem como base o esquema Dublin Core[89]. Dispõe, em sua versão de instalação, dois esquemas, Dublin Core Qualificado [90] e Dublin Core Terms[91], que podem ser estendidos, como realizado no repositório DataShare[92].

Com relação aos metadados técnicos [DOC5], na submissão de um arquivo que compõem um item, DSpace identifica o formato do arquivo submetido (pela terminação indicada no nome do arquivo) e registra essa informação através do metadado de Dublin Core, **dc.format**. A ferramenta também registra outras informações a respeito do arquivo (descrição, e checksum), mas essas informações ficam armazenadas no banco de dados relacional que dá suporte à ferramenta. DSpace não extrai automaticamente e nem registra metadados técnicos a partir dos arquivos submetidos.

Com relação a metadados de proveniência [DOC1], DSpace possui um fluxo para submissão de item e possibilita também a submissão por máquina. As ações realizadas na execução da submissão são armazenadas como metadados de proveniência do item submetido (elemento **dc.description.provenance**). A Figura 8 apresenta os metadados que registram a submissão, a aprovação e a publicação de um item. Esses metadados registram pessoa envolvida, a data/hora da operação, e o nome, o tamanho e o código *hash* (**MD5**) do arquivo (para fixidez).

Informações adicionais de proveniência, como contexto de produção e recursos relacionados, referentes às atividades realizadas para produzir os dados e aos agentes envolvidos[93] podem ser registradas através da criação de novos metadados para esse fim[94] e da configuração da interface de submissão, para permitir que esses metadados sejam informados pelos produtores.

89 Esquema de Metadados Dublin Core - http://dublincore.org/documents/dces/

90 Esquema Dublin Core Qualificado por DSpace – Dspace6 Documentation https://wiki.duraspace.org/display/DSDOC6x/Metadata+and+Bitstream+Format+Registries

91 Esquema Dublin Core Terms - http://dublincore.org/documents/dcmi-terms/

92 Metadados do Repositório DataShare - https://www.wiki.ed.ac.uk/display/datashare/metadata

93 Ontologia de Proveniência – PROV https://www.w3.org/TR/2013/REC-prov-o-20130430/

94 Mapeamento PROV-Dublin Core - https://www.w3.org/TR/2013/NOTE-prov-dc-20130430/

Com relação ao registro de ações de preservação digital [DOC2], na submissão via interface de usuário, DSpace conduz os usuários na criação e na avaliação do pacote de submissão, registrando, na forma de metadados de proveniência, o resultado dessas ações. DSpace possui algumas funções de curadoria digital (como verificação de ocorrência de vírus e da falta de metadados obrigatórios[95]), mas os resultados dessas ações não são registrados em metadados.

Com relação a metadados estruturais, a descrição do pacote de informação [DOC4], 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. Com isso, DSpace não armazena um único local (sistema de arquivos), de forma independente do *software*, todas as informações que compõem o pacote do item, incluindo arquivos e metadados de todos os tipos. Isso traz dificuldades para a preservação digital.

Entretanto, DSpace permite a exportação de pacotes, e essa exportação segue padrões de estruturas e metadados recomendados para preservação digital [83]. Nos pacotes exportados por DSpace, os metadados estruturais são representados no padrão METS[96], a partir das informações de estrutura que estão na base de dados relacional. Outros metadados administrativos e técnicos também são codificados a partir de informações da base de dados, nos padrões METS e PREMIS[97], incluindo informações de coleções, grupos de usuários. Essa funcionalidade é recomendada para dar apoio a estratégias de cópia de segurança e de preservação digital.

Com relação ao uso de vocabulários controlados, DSpace não dispõe de interface para criação e gestão de vocabulários controlados. Entretanto, dentre os recursos disponíveis por DSpace para a definição de formulários de entradas de dados, a ferramenta possibilita que dados sejam controlados por vocabulários fornecidos em arquivos XML[98] [DOC6].

DSpace dispõe também de um recurso que usa indexador Solr para controlar autoridade[94]. Os nomes de autoridade a serem usados no preenchimento de metadados de autoridades são gerenciados através da ferramenta de indexação Solr, por meio de um índice de autoridades. 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 ORCID[99].

95 DSpace 6.x Documentation - Curation System - https://wiki.duraspace.org/display/DSDOC6x/Curation+System

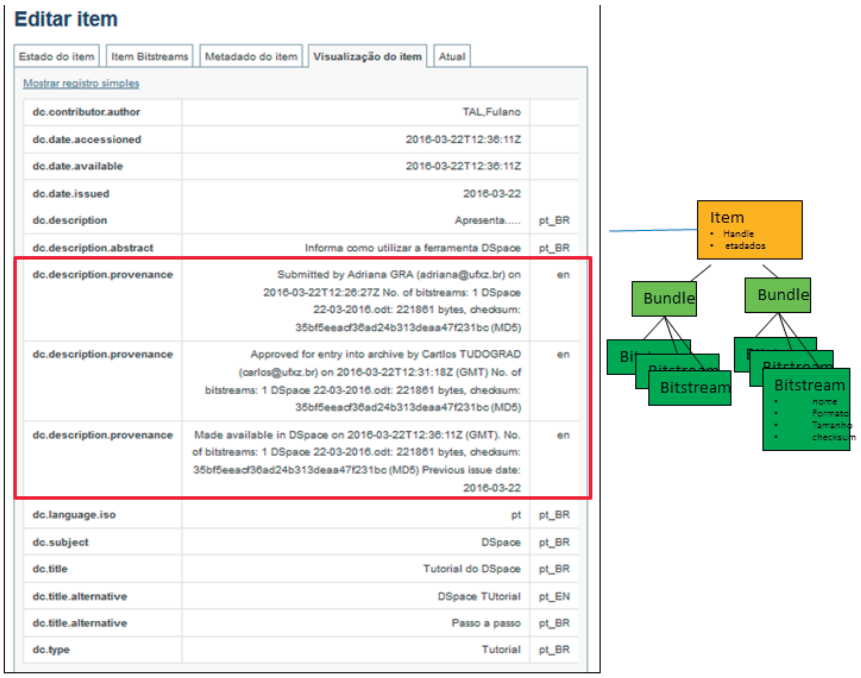
96 METS - Esquema para Metadados Estruturais - Metadata Encoding and Transmission Standard - http://www.loc.gov/standards/mets/

97 PREMIS - Esquema para Medadados de Preservação Digital Eshttps://www.loc.gov/standards/premis/

98 DSpace 6.x Documentation - Configuring Controlled Vocabularies https://wiki.duraspace.org/display/DSDOC6x/Submission+User+Interface#SubmissionUserInterface-ConfiguringControlledVocabularies

99 DSpace 6.x Documentation - IRCID Integration - https://wiki.duraspace.org/display/DSDOC6x/ORCID+Integration

Figura 8 - Metadados de Proveniência em DSpace



Fonte: Dados da pesquisa

DSpace oferece recursos para especificar mapeamento entre esquemas de metadados (*crosswalks*), para permitir a colheita de metadados em diversos esquemas via protocolo OAI-PMH. O mapeamento é realizado através de regras XSLT[100] [DOC8]. DSpace, em sua instalação padrão, já disponibiliza regras XSLT para permitir a colheita de metadados em esquemas como RDF, METS, MODS[101] e MARC[102].

DSpace permite a colheita de metadados em documento no formato RDF/XML, padrão da Web Semântica, embora não permita o armazenamento e manipulação de metadados na forma de triplas RDF [DOC9]. DSpace-CRIS é uma extensão de DSpace que permite a descrição de recursos de Sistemas de Informação de Pesquisa [DOC10].

DSpace não dispõe de recursos específicos para gerenciar a documentação que dá apoio ao uso de dados (como codebooks/livros de códigos). Entretanto, é possível configurar uma estrutura de item (via *bundle* e *bitstreams*) que representa arquivos de documentação e uma interface de submissão que conduzem usuários a incluírem adequadamente esses arquivos junto ao item. Também é possível registrar essas informações em metadados específicos [DOC11].

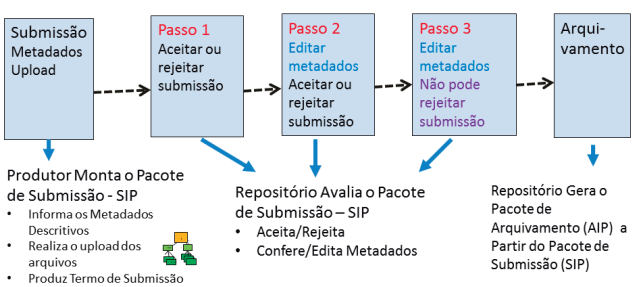
100 DSpace 6.x Documentation - Client-side stylesheet - https://wiki.duraspace.org/display/DSDOC6x/OAI+2.0+Server
101 Metadata Object Description Schema (MODS) - http://www.loc.gov/standards/mods/
102 MARC - https://www.loc.gov/marc/

4.1.5 Produção dos Conjuntos de Dados

Essa seção analisa DSpace com relação aos critérios: gerenciamento de acordos de submissão e licença [SUB1], fluxos de submissão [SUB2], transferência legal da custódia dos dados ao repositório [SUB3], validação e verificação pacotes de submissão [SUB4], extração, verificação e/ou produção de metadados técnicos, administrativos e descritivos [SUB5], submissão em lote, por máquina e em ambientes distribuídos [SUB6], uso de padrões [SUB7].

DSpace permite a definição de acordos de submissão específicos para cada coleção, assim como disponibiliza aos usuários consumidores dos itens a informação da licença de uso estabelecida nesse acordo. Não gerencia, entretanto, licenças, isto é, não possui um módulo em que as licenças e os acordos são cadastrados e relacionados com os itens aos quais dizem respeito [SUB1].

Figura 9 - Fluxo de Submissão



Fonte: Dados da pesquisa

DSpace permite a configuração de um fluxo de submissão para cada coleção [SUB2]. Disponibiliza em sua instalação padrão quatro etapas de fluxo, que podem ser usadas na definição de um fluxo de submissão a uma coleção (Figura 9). Através 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, devidamente autorizado e autenticado, assinala que concorda com os termos de submissão, que é então adicionado ao pacote.

Através 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, DSpace permite a transferência de custódia do objeto submetido ao repositório [SUB3].

A validação e a verificação de um pacote submetido via interface de usuário é realizada através dos seguintes mecanismos: DSpace disponibiliza uma interface para

construção do pacote de submissão, direcionado o usuário a produzir um pacote com relação a sua estrutura adequada. Então, etapas do fluxo podem ser criadas, através das quais pessoas irão aprovar a validade do item, com os resultados sendo registrados em metadados de proveniência [SUB4]. Dspace não possui microserviços que verificam e validam automaticamente pacotes submetidos através desse fluxo.

DSpace não dispõe de microserviços que extraem automaticamente metadados técnicos dos arquivos submetidos (como informações sobre formato, versões de formatos, configurações de imagens) e que validam formatos[103] [SUB4] [SUB5]. Possui o recurso "Curation Task", que permite o desenvolvimento de novas tarefas (via extensão de classes do *software*) e que podem ser chamadas no *workflow* de submissão. Disponibiliza tarefas que permitem buscar metadados em outras bases de dados[104] e usá-los na descrição do item.

DSpace permite a submissão por máquinas, via rede e/ou em lote [SUB6]. Itens podem ser submetidos em lote, desde que suas estruturas estejam de acordo com estruturas suportadas para pacotes (DSpace Archival Information Package e DSpace METS SIP[105]) . Através do protocolo Sword[54], outros ambientes (sistemas computacionais) podem submeter itens ao repositório. DSpace também permite que coleções sejam alimentadas a partir de metadados/itens que estão em outros repositórios, via protocolos OAI-PMH[55] e OAI-ORE [56] [SUB7]. Isso permite o uso de DSpace como agregador de informações em um ambiente federado de repositórios. [SUB6]

4.1.6 Armazenamento a Longo Prazo e Planejamento da Preservação

Essa seção analisa DSpace com relação aos critérios: serviços/microserviços para garantir acesso a longo prazo [PD1], planejamento e ações de preservação digital [PD2], integração com serviços de preservação digital de terceiros ou colaborativos [PD3], exclusão de dados [PD4] e uso em Repositório Digital Confiável certificado [PD5].

DSpace possui o um módulo de curadoria digital[106] que permite o desenvolvimento de novas tarefas (microserviços) de curadoria, a partir da extensão de classes do *software*. Essas tarefas podem ser configuradas para serem chamadas pelo fluxo de submissão ou pelos administradores do repositório. A ferramenta já possui desenvolvidos alguns microserviços de curadoria digital, que permitem verificar a integridades dos *links* dos objetos, checar a ocorrência de vírus, checar se campos obrigatórios estão presentes. DSpace, entretanto não dispõe, na sua distribuição de instalação, de microserviços que identificam e validam formatos[98], usando por exemplo ferramentas como JHOVE [61] ou DROID [60] [PD1].

103 DSpace – Development – Format Identification Issues - https://wiki.duraspace.org/display/DSPACE/FormatIdentificationIssues

104 DSpace 6.x Documentation - Start Submission Lookup Step https://wiki.duraspace.org/display/DSDOC6x/Submission+User+Interface#SubmissionUserInterface-ConfiguringStartSubmissionLookupStep

105 DSpace 6.x Documentation - Importing and Exporting Content via Packages https://wiki.duraspace.org/display/DSDOC6x/Importing+and+Exporting+Content+via+Packages

106 DSpace 6.x Documentation - https://wiki.duraspace.org/display/DSDOC6x/Curation+System https://wiki.duraspace.org/display/DSDOC6x/Curation+System

Um importante recurso para apoio à preservação digital é a funcionalidade de DSpace que permite a exportação/importação dos pacotes de armazenamento de informação em formato que segue recomendações de preservação digital[85]. Isso permite que ações de refrescamento, replicação e até mesmo migração possam ser realizadas tendo como fonte esses pacotes. Nesse caso, a preservação seria realizada com base nesses pacotes, e DSpace seria o mecanismo de acesso e ingestão dos itens [PD1].

Com relação a funcionalidades que apoiam o planejamento e ações de preservação digital [PD2], DSpace controla formatos, através do registro de formatos, no qual é possível especificar quais formatos são aceitos, quais são os formatos que repositório irá se comprometer com a preservação, e quais formatos não são aceitos (Registro de Formatos - Figura 4). Não disponibiliza recursos em que são executadas ações automáticas que gerenciam formatos com relação a políticas de formatos, como converter automaticamente arquivos submetidos em formatos não aceitos para arquivos em formatos aceitos [PD2]. Cabe novamente ressaltar que DSpace identifica o formato de um arquivo pela terminação presente no nome do arquivo, isto é, não usa ferramentas como JHOVE e DROID para identificar os formatos dos arquivos a partir da análise de suas estruturas.

DSpace permite a participação do repositório em redes cooperadas de preservação digital que usam o protocolo Lockss[59], em que os dados são replicados nos repositórios membros, possibilitando a reconstrução dos dados de um repositório em caso de perda. A Rede Cariniana[107] é uma iniciativa brasileira, promovida pelo IBICT, no qual fazem parte repositórios nacionais que usam DSpace, como o Lume[108], da UFRGS [PD3]. DSpace também pode ser integrado com Arquivematica[62]. Nesse caso, pacotes exportados do Dspace podem ser transferidos para Arquivematica, que assume a função preservação a longo prazo para esses pacotes[109] [PD3].

DSpace permite que dados sejam embargados e removidos [PD4]. Quando um item é removido, ele ainda fica disponível aos administradores gerais do DSpace, podendo ser restaurado. Um administrador geral pode remover um item em caráter permanentemente.

Repositórios digitais que usam DSpace obtiveram certificações Data Seal Approval[110] e Core Trust Seal[111], como Dryad[84] e Datashare[79], demonstrando que a ferramenta não traz impedimentos para tal [PD5].

107 Rede Cariniana - http://cariniana.ibict.br/

108 Lume – Repositório Institucional da UFRGS - https://lume.ufrgs.br/

109 Transferência para Archivematica via Dspace Export - https://www.archivematica.org/en/docs/archivematica-1.7/user-manual/transfer/dspace/#dspace

110 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

111 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

4.1.7 Acesso e Uso dos Conjuntos de Dados

Essa seção analisa DSpace com relação aos cvritérios: recuperação de informação [AC1], informações sobre direitos de uso, licenças [AC2], informações de proveniência e para uso dos dados [AC3], informações de citação [AC4], identificadores globais e persistentes [AC5], identificadores globais e persistentes para recursos relacionados [AC6], acesso aos dados e aos metadados via identificador e protocolo aberto [AC7], restrições de acesso [AC8], Gerenciamento e autenticação de usuários envolvidos com acesso [AC9], entrega de dados ao consumidor [AC10], entrega dos metadados ao consumidor [Ac11], ferramentas para visualização e análise de dados [AC12], estatísticas e relatórios de uso [AC13], acesso às descrições em formatos para Linked Data/Web Semântica [AC14], recuperação em ambiente Linked Data/Web Semântica [AC15].

DSpace permite a busca por expressões e navegação, através da configuração de quais metadados serão disponibilizados para essas funções. Ordena resultados por estimativa de importância [AC1] e apresenta informações aos usuários sobre a licença de uso [AC2].

Ações que envolveram a submissão são registradas pelo DSpace no metadado **dc.provenance**. Em sua instalação padrão, esses metadados não são configurados para serem acessíveis pelos usuários consumidores. Como já mencionado, DSpace permite a criação de novos metadados para registrar informações de proveniência (com processo e atores envolvidos na produção dos dados) e para uso dos dados, que podem ser disponibilizados aos consumidores. Também permite o armazenamento junto ao pacote do conjunto de dados de documentação para uso dos dados (como livros de códigos/codebooks) [AC3].

DSpace permite a configuração da interface de exibição de cada item. Esses recursos podem ser usados para gerar informações de citação a partir dos metadados de citação [AC4]. A Figura 10 e a Figura 11 apresentam, respectivamente, informações de citação nos repositórios Datashare e Dryad.

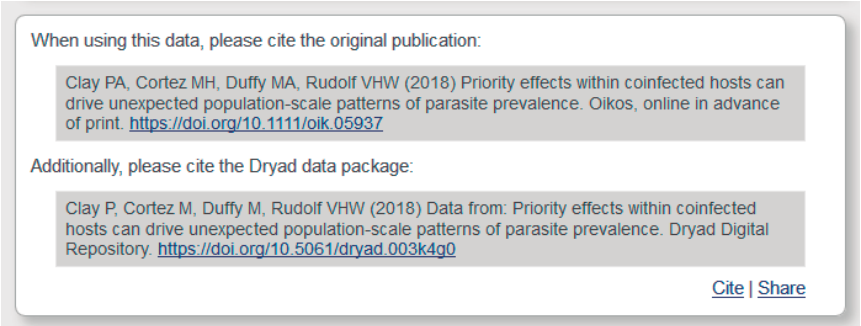
Figura 10 - Citação em Datashare

Citation

Echavarri-Bravo, Virginia; Tinzl, Matthias; Cruickshank, Faye; Mackay, C Logan; Clarke, David. (2018). Characterisation of enzymatic processing of commercial (Kraft) lignin by high resolution FT-ICR mass spectrometry – Dataset., [dataset]. University of Edinburgh, School of Chemistry. http://dx.doi.org/10.7488/ds/2459.

Fonte: Repositório DataShare - https://datashare.is.ed.ac.uk/handle/10283/3205

Figura 11 - Citação em Dryad



Fonte: Repositório Dryad - https://datadryad.org/resource/doi:10.5061/dryad.003k4g0

Em DSpace, objetos e metadados são acessados via serviços de identificadores globais e via protocolos abertos: usa os serviços DOI [63] e HandleSystem [b] para identificação, e HTTP e HTTPS para acesso e transferência de informações entre o cliente e o repositório. [AC5] [AC7]. DSpace foi concebido e usa o serviço Handle System para prover identificadores globais e persistentes para seus itens [AC5]. Também permite o uso de DOI. O uso de DOI implica em registrar cada identificador DOI em uma agência de registro DOI, como DataCite[112] ou EZID[113]. DSpace permite que DOIs sejam registrados em DataCite ou EZID, através de funções que chamam as APIs de registros dessas agências[114]. A Figura 12 mostra um conjunto de dados identificado por DOI e HandleSystem, armazenado no repositório da Universidade de Minesota (DRUM). A Figura 13 mostra os metadados desse conjunto de dados registrados por Dspace em DataCite, associados ao DOI do objeto.

DSpace permite o uso do serviço ORCID para controle dos nomes de autoridade, no preenchimento de metadados que representam autoridades. 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[115] [AC6].

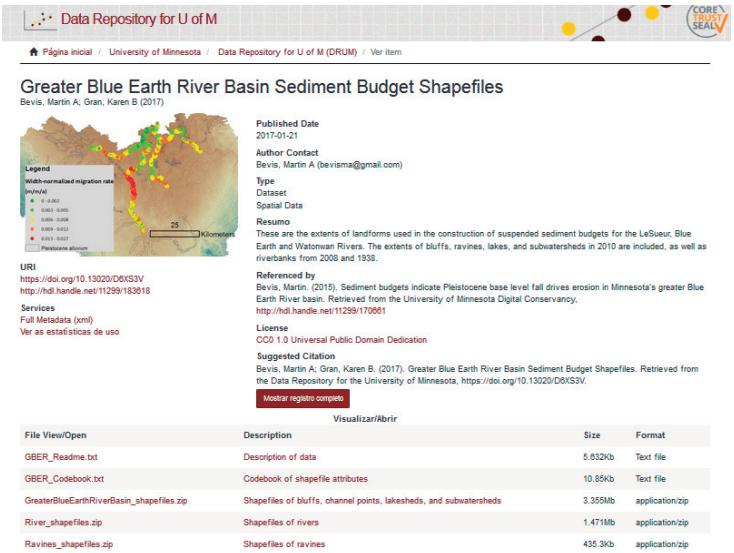
112 DataCite – Atribuindo DOI - https://www.datacite.org/dois.html

113 EZID – DOI - https://ezid.cdlib.org/

114 DSpace 6.x Documentation DOI - https://wiki.duraspace.org/display/DSDOC6x/DOI+Digital+Object+Identifier

115 DSpace 6.x Documentation - ORCID Integration - https://wiki.duraspace.org/display/DSDOC6x/ORCID+Integration

Figura 12 - DOI no Repositório DRUM - Universidade de Minesota



Fonte: Repositório DRUM - https://doi.org/10.13020/D6XS3V

Permite acesso aos dados e aos metadados via identificador (Handle System ou DOI) e protocolo aberto (Http, Https) [AC6]. DSpace possibilita a definição de políticas de acesso para comunidades (incluindo suas subcomunidades e coleções), coleções e item, através da criação de grupos de usuários e da atribuição de autorizações/restrições de acesso a esses grupos. Permite também restrição de acesso por motivos de embargo [AC8].

Figura 13 - Metadados em Datacite do objeto com DOI 10.13020/d6xs3v


```
<?xml version="1.0" encoding="UTF-8"?>
- <resource xsi:schemaLocation="http://datacite.org/schema/kernel-3
http://schema.datacite.org/meta/kernel-3/metadata.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns="http://datacite.org/schema/kernel-3">
    <identifier identifierType="DOI">10.13020/D6XS3V</identifier>
  - <creators>
    - <creator>
        <creatorName>Bevis, Martin A.</creatorName>
      </creator>
    - <creator>
        <creatorName>Gran, Karen B.</creatorName>
      </creator>
    </creators>
  - <titles>
      <title>Greater Blue Earth River Basin sediment budget shapefiles</title>
    </titles>
    <publisher>Data Repository for the University of Minnesota (DRUM)</publisher>
    <publicationYear>2017</publicationYear>
    <resourceType resourceTypeGeneral="Dataset">Dataset</resourceType>
  - <rightsList>
      <rights rightsURI="https://creativecommons.org/publicdomain/zero/1.0/">CC0 1.0 Universal</rights>
    </rightsList>
</resource>
```


Fonte: Datacite - https://data.datacite.org/application/vnd.datacite.datacite+xml/10.13020/d6xs3v

Os usuários envolvidos com o acesso podem ser autenticados pelo próprio do DSpace e também pelo uso dos serviços LDAP, Shibboleth, IP Address e certificado X.509. A autenticação via Shibboleth permite que usuários sejam automaticamente inseridos como membros de grupos de usuários, no momento da autenticação, tendo como base informações do usuário disponibilizadas no serviço Shibboleth. Dessa

forma, por exemplo, a partir de informações de Shibboleth sobre a afiliação de um usuário autenticado, DSpace pode colocar este usuário no grupo que representa essa instituição, fazendo que este usuário receba todas as autorizações dadas a este grupo. [AC9]. Objetos podem possuir diferentes representações, isto é, um objeto pode estar representado em mais de um formato de arquivo, possibilitando a entrega em variadas representações, a fim de atender variados consumidores. Além disso, outros sistemas informatizados podem obter objetos do DSpace, através do protocolo OAI-ORE [AC10]. Metadados podem ser obtidos em diversos esquemas, através do protocolo de colheita OAI-PMH [AC11].

DSpace oferece uma interface configurável que traz facilidade para a sua integração com visualizadores de objetos [AC12]. 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. Uma série de outras estatísticas específicas são computadas. Por exemplo, o *software* Solr fornece quantidade de visitas às *homepages* de comunidades, coleções e itens bem como os dez países e dez cidades com mais visitantes [AC13].

Dspace permite a colheita de metadados em documento RDF/XML, padrão da Web Semântica [AC14]. Não permite o armazenamento e manipulação de metadados na forma de triplas RDF [AC15], mas possibilita sua exportação para bases de dados de triplas. DSpace CRIS é uma extensão de DSpace que permite a descrição de recursos de Sistemas de Informação de Pesquisa .

4.2 Dataverse

Esta seção detalha o estudo da solução tecnológica Dataverse.

4.2.1 Desenvolvimento, Manutenção e Uso do *Software*

Essa seção analisa Dataverse com relação aos critérios: tecnologia e plataforma [SW1], distribuição e versionamento [SW2], estratégia de desenvolvimento do *software* [SW3], licença de uso [SW4], desempenho e escalabilidade [SW5], Presença, usuários e uso no Brasil [SW6].

Dataverse vem sendo desenvolvido desde 2006 no Harvard's Institute for Quantitative Social Science (IQSS), junto de muitos colaboradores em todo o mundo. Derivou do projeto Virtual Data Center, uma colaboração entre o Harvard-MIT Data Center e a Harvard University Library.

Hoje, o *software* Dataverse é disponibilizado oficialmente apenas para o sistema operacional Linux [116] [SW1]. Entre os principais requisitos de plataforma de *software* destacam-se:

116 Dataverse - https://github.com/IQSS/dataverse/releases

- Java Oracle JDK 8[117]
- Glassfish 4.1[118]
- PostgreSQL[119] 9.6 (with pgcrypto)
- Apache Solr 7.3[120]
- jq 1.4+ [121]
- ImageMagick[122]
- R 3.5[123]
- Rserve[124]
- Servidor SMTP
- DOI[125] ou Handle[126]

A configuração de *hardware* mínima recomendada para um sistema em produção inclui pelo menos 8GB de memória RAM e 50 GB de espaço em disco.

O esquema de distribuição e versionamento do *software* Dataverse [SW2] inclui versões principais (*major releases*) com novas funcionalidades, mudanças arquiteturais ou de sistema. Estas são caracterizadas pela numeração [x].[y]. Versões menores (*minor releases*) são geradas a partir de correção de falhas (*bug fixes*) em versões principais, sendo definidas por [x].[y].[z]. Atualmente, a última versão é 4.9.3.

Dataverse é distribuído sob licença Apache 2.0[127], permitindo ao usuário a liberdade de usar o *software* para qualquer finalidade, distribuí-lo, modificá-lo e distribuir versões modificadas, sob os termos da licença, sem preocupação com *royalties* [SW4].

Características de desempenho e escalabilidade do Dataverse dependem diretamente da configuração do servidor Java web, do sistema gerenciador de banco de dados relacional e do sistema de recuperação de informações [SW5]. PostgreSQL deve ser sintonizado para admitir *buffers* compartilhados de tamanho adequado, *checkpoints* de escrita dos *logs* em momentos oportunos, entre outras várias configurações. Testes de carga e de desempenho são disponibilizados em um conjunto de *scripts*[128] usando o *framework* Locust[129].

Por fim, existem apenas duas instalações no Brasil registradas na comunidade Dataverse[130]: a Rede Cariniana do IBICT e o Repositório de Dados de Pesquisa da UFABC. Na comunidade Re3data[131] está registrada apenas a instalação do IBICT [SW6].

117 Oracle JDK https://www.oracle.com/technetwork/java/javase/downloads/index.html
118 GlassFish https://javaee.github.io/glassfish/download
119 PostgreSQL http://www.postgresql.org/
120 Apache Solr http://lucene.apache.org/solr/
121 jq https://stedolan.github.io/jq/
122 ImageMagick https://www.imagemagick.org/
123 R https://www.r-project.org/
124 Rserve https://rforge.net/Rserve/
125 DOI http://www.doi.org/
126 Handle https://www.handle.net/
127 Licança Dataverse - https://github.com/IQSS/dataverse/blob/develop/LICENSE.md
128 Scripts de Teste de Dataverse - https://github.com/IQSS/dataverse-helper-scripts/tree/master/src/stress_tests
129 Locust https://locust.io/
130 Comunidade Dataverse - https://dataverse.org/
131 Diretório de Repositórios Re3Data - https://www.re3data.org/browse

4.2.2 Representação do Ambiente do Repositório

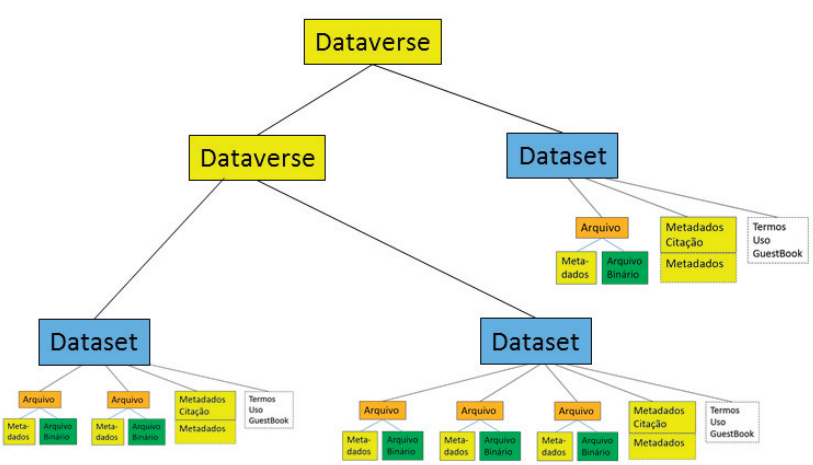
Essa seção analisa Dataverse com relação aos critérios: representação do ambiente [AMB1], recursos para operacionalizar políticas de funcionamento do ambiente [AMB2], recursos para estabelecer políticas descentralizadas e autônomas [AMB3], representação de ambiente integrada com Sistemas de Informação de Pesquisa [AMB4], representação de ambiente para Web Semântica / Dados Abertos e Ligados [AMB5], recursos que permitam transparência e feedback aos envolvidos [AMB6].

Dataverse foi desenvolvido para atender a demandas de repositórios para dados de pesquisa. Possui recursos que permitem variadas configurações para ambientes de repositório de dados [AMB1].

> 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 para pesquisadores e editores de dados e incentivá-los a compartilhar seus dados. Ao instalar o *software* Dataverse Network, uma instituição pode hospedar vários arquivos virtuais individuais, chamados "dataverses" para acadêmicos, grupos de pesquisa ou periódicos, fornecendo uma estrutura de publicação de dados que suporta reconhecimento de autor, citação persistente, descoberta e preservação de dados[132].

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, tendo em vista que dataverses podem conter outros dataverses. Cada dataverse contém **datasets**, que são as entidades que representam conjuntos de dados, na forma de estudos. [AMB1] (Figura 14).

Figura 14 - Dataverses e Datasets

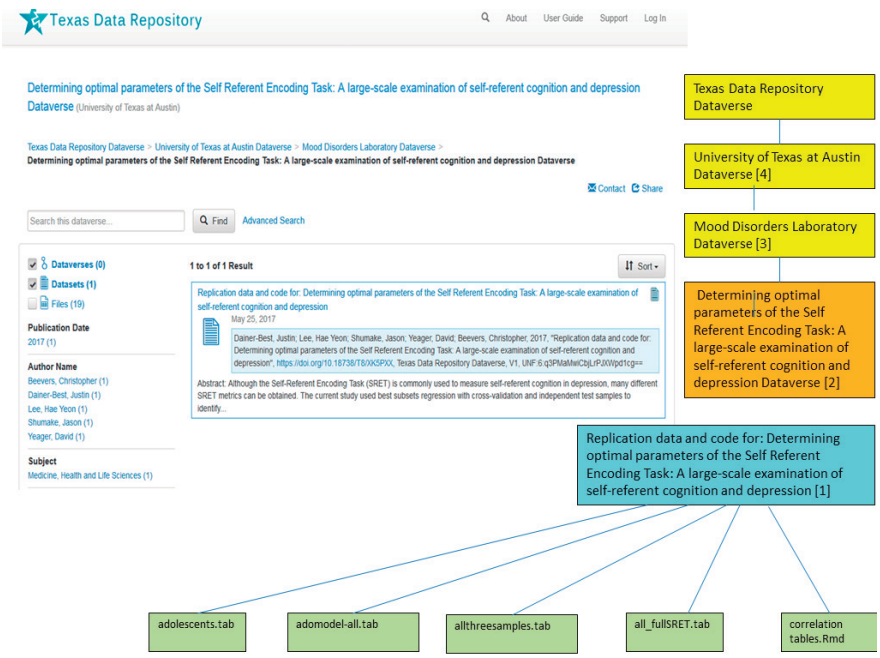


Fonte: Dados da pesquisa

132 Mercè Crosas. The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. DLIB Magazine, 2011 - http://dlib.org/dlib/january11/crosas/01crosas.html

A Figura 15 exemplifica o uso do *software* Dataverse pelo Repositório Texas Digital Data (TDL). TDL é um repositório multi-institucional, que armazena dados de várias universidades do Texas. Nesse exemplo, observamos um estudo [1-na figura] (dataset), que faz parte de um conjunto de estudos [2] (dataverse), que pertence a um Laboratório [3] (dataverse) da Universidade do Texas em Austin [4] (dataverse).

Figura 15 - Texas Data Repository



Fonte: Elaborado pelos autores a partir do Repositório TDL - https://dataverse.tdl.org/dataverse/sretdepression

O *software* dispõe de recursos para implementar dataverses com políticas de funcionamento próprias e distintas [AMB2][AMB3].

> A entidade dataverse permite que seja dada a autores a propriedade de arquivos de dados na web. Cada dataverse fornece citação acadêmica, marca personalizada (interface), descoberta de dados, controle sobre atualizações e termos de acesso e uso [128].

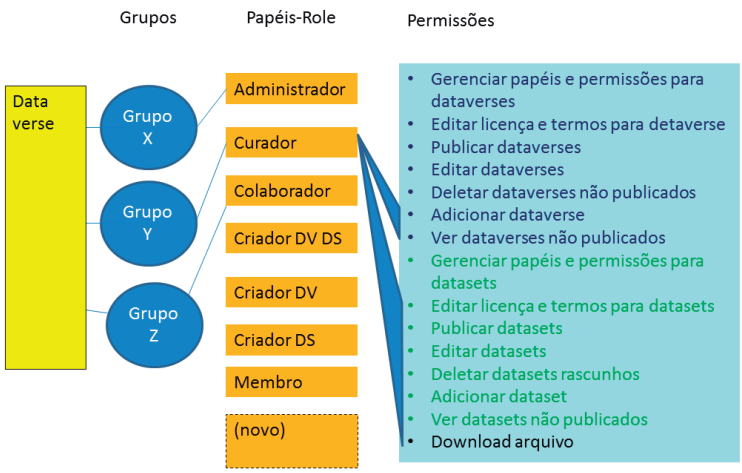
No *software* Dataverse, uma entidade dataverse pode ser configurada como se fosse um repositório independente, com todas autorizações para gerenciamento e operação (como definir papéis, permissões e grupos; criar e gerenciar dataverses, datasets e arquivos, etc.) [AMB3]. Para a definição de políticas, o *software* trabalha os seguintes conceitos [AMB2]:

- **Permissão**: representa a permissão para realização de uma determinada funcionalidade do ambiente, como publicar dataverse, editar dataverse, gerenciar papéis e permissões de dataverse ou datasets.

- **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 (como Administrador, Publicador de Dataverse, Curador de Dataverse e Editor de Dataset, etc.), e novos papéis podem ser criados.
- **Grupo**: representa um grupo de usuários, que são habilitados a atuar com determinados papéis.

A Figura 16 apresenta as permissões do sistema, indicando (relacionando) aquelas que estão autorizadas às pessoas que possuem o papel de Curador. Os usuários do grupo Y são aqueles que possuem o papel de Curador [AMB3].

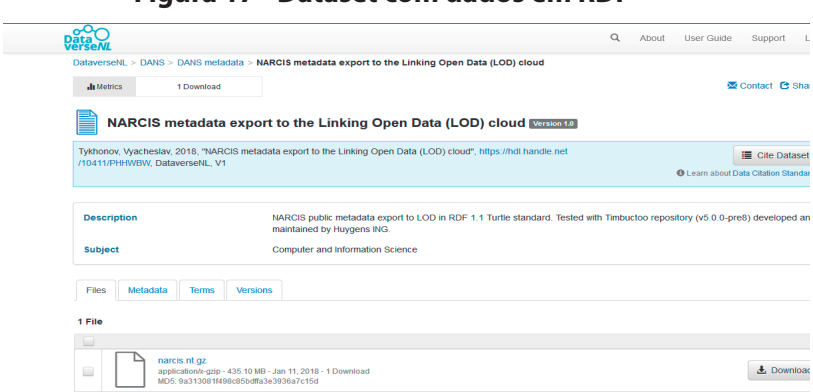
Figura 16 - Grupos, Papéis e Permissões



Fonte: Dados da pesquisa

Na distribuição do *software* Dataverse não estão disponíveis funcionalidades que buscam integração com Sistemas de Informação de Pesquisa [AMB4], assim como armazenamento de metadados na forma de triplas-RDF, recuperáveis através da linguagem SPARQL [AMB5]. Dataverse, entretanto, permite a disponibilização de arquivos XML que contêm sentenças RDF, como exemplificado na Figura 17.

Figura 17 - Dataset com dados em RDF



Fonte: Dataverse NL - https://hdl.handle.net/10411/PHHWBW

O Dataverse dispõe dos seguintes recursos para prover transparência e *feedback* aos envolvidos no processo de submissão de um conjunto de dados, que é realizado por meio de fluxos bem definidos, por pessoas autorizadas, com rastreabilidade e comunicação com os envolvidos [AMB6]:

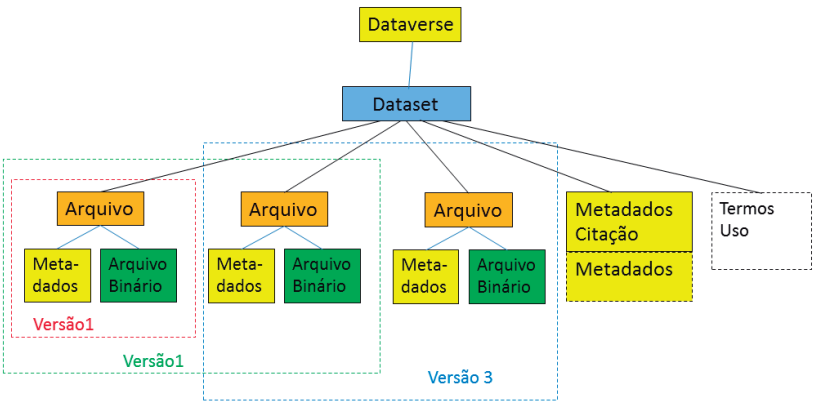
- Políticas de autorizações (via permissões, papéis e grupos - Figura 16) são atribuídas a entidades dataverses e a datasets, identificando aqueles que irão exercer papéis que permitem administrar, criar e editar datasets e dataverses, editar licenças, publicar dataverses etc.
- Um dataset, quando criado, assume o estado de versão "DRAFT", isto é, não publicada. Na criação, o usuário produz os metadados, carrega os arquivos e edita as licenças. O ambiente pode ser configurado para permitir a definição de uma política em que usuários de determinado grupo estão autorizados a criar um dataset, usuários de um outro grupo estão autorizados a estabelecer a licença do dataset, e usuários de ou terceiro grupo podem publicar o dataset, por exemplo.
- O ambiente possibilita que um dataset criado seja submetido para revisão, para ser então publicado. O ambiente notifica a solicitação de revisão aos curadores, que decidem então se o dataset será publicado ou se retornará aos autores, com a notificação da decisão. Sendo publicado, o dataset assume a Versão 1, e a informação de quem a publicou a versão é registrada.
- Quando um dataset é atualizado, pela edição de metadados ou remoção/adição de arquivos, uma nova versão deste dataset é criada. Essa versão inicialmente assume o estado "DRAFT" e é de acesso restrito. Quando publicada por pessoa devidamente autorizada, torna-se então uma nova versão do dataset, assumindo uma numeração de versão.
- O ambiente notifica os envolvidos. Cada usuário, em seu perfil, tem acesso a todas as notificações a ele direcionadas, como, por exemplo, a uma notificação de solicitação de revisão de um dataset no qual exerce o papel de curador.

4.2.3 Representação dos Conjuntos de Dados

Essa seção analisa Dataverse com relação aos critérios: natureza dos conjuntos de dados [PAC1], estruturas para representar os conjuntos de dados [PAC2], formatos de arquivos [PAC3], recursos para versionamento de conjuntos de dados [PAC4], dados cinco estrelas [PAC5], uso de padrões da web semântica/linked data [PAC6].

Dataset (Figura 18) é o recurso que o ambiente disponibiliza para armazenar um conjunto de dados, considerando que o mesmo é um resultado de um estudo. Um dataset é composto por metadados, pelos termos de uso (como licenças) e por arquivos. Metadados descrevem o dataset e os arquivos.

Figura 18 - Dataset



Fonte: Dados da pesquisa

A estruturas de um dataset são adequadas para representar dados de pesquisa, sendo compatíveis com padrão de metadados DDI Lite48[133] e DDI 2.5 Codebook[134]. Esses esquemas de metadados visam descrever dados de pesquisa no contexto de um estudo e envolvem tanto metadados descritivos, quando metadados estruturais, que descrevem os arquivos que compõe o pacote [PAC2]. No caso de conjuntos de dados tabulares, ocorre também a descrição das estruturas das variáveis [PAC2]. No *software* Dataverse, grupo de pesquisa/unidade, estudo e conjunto de dados podem ser representados da seguinte forma:

- **Unidade ou Grande Grupo**: Dataverse
- **Grupo de pesquisa**: Dataverse
- **Estudo**: Dataset
- **Conjunto de Dados do Estudo**: Arquivos

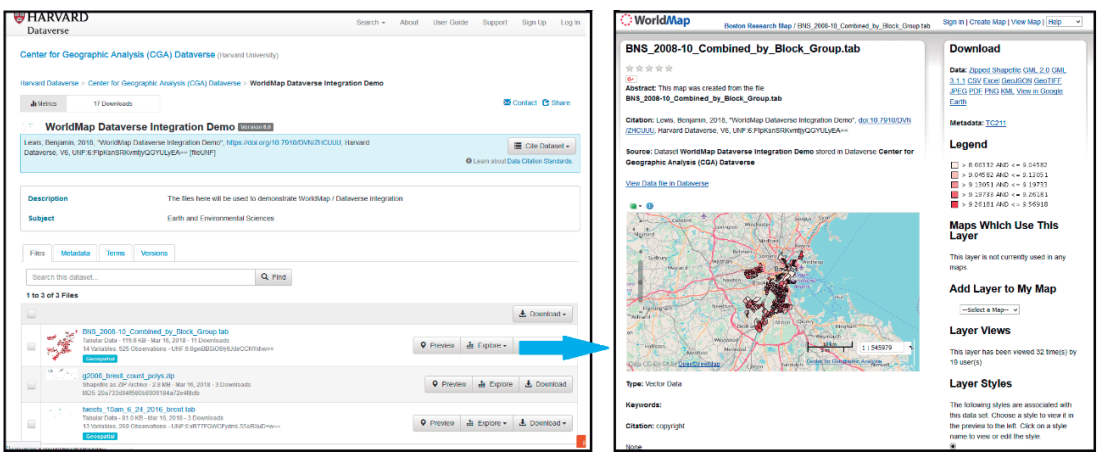
O ambiente aceita arquivos de diversos formatos [PAC1] e utiliza a ferramenta JHOVE para identificar, validar e descrever o formato de um arquivo submetido [PAC3]. Entretanto, não disponibiliza funcionalidades para gestão de formatos, isto é, para permitir o estabelecimento de políticas e de controle de formatos aceitos, por exemplo [PAC3].

O ambiente possui recursos especiais para tratar arquivos em formatos tabulares (planilhas e arquivos de aplicativos de análise estatística, com o SPSS e R) e geoespaciais. Com relação a arquivos com dados tabulares submetidos (formatos CSV, Excell, SPSS, R, etc.), Dataverse gera representações para esses arquivos em um formato canônico (.tab), para permitir que arquivos de dados tabulares possam ser usados uniformemente por ferramentas externas (como R) e por ferramentas já integradas com o ambiente, como TwoRavens[135] (dados tabulares) e WordMap[136] (dados geoespaciais) [PAC1].

133 DDI Lite – Elementos recomendados de DDI Codebook http://www.ddialliance.org/sites/default/files/ddi-lite.html
134 DDI Codebook 2.5 - http://www.ddialliance.org/Specification/DDI-Codebook/2.5/
135 TwoRavens - http://2ra.vn/
136 Worldmap - http://worldmap.harvard.edu/

Dataverse verifica (confere a integridade) e armazena dados geoespaciais em formatos vetoriais (*shapefiles*), a fim de permitir sua integração com a ferramenta WordMap. Também permite que o produtor, através de uma ferramenta interativa, estabeleça o mapeamento de dados tabulares que contêm informações geoespaciais para WordMap. A Figura 19 apresenta o dataset com DOI 10.7910/DVN/ZHCUUU, que exemplifica o mapeamento e a visualização de dados tabulares em WorldMap.

Figura 19 - Dados Tabulares com Informações Geoespaciais em Dataverse/

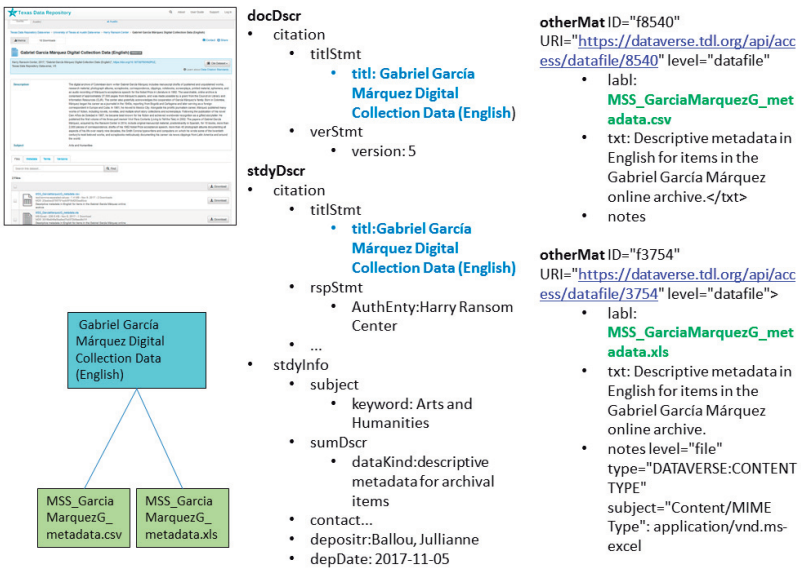


Fonte: Repositório Harvard Dataverse - https://doi.org/10.7910/DVN/ZHCUUU

O ambiente permite o versionamento de datasets. Cada alteração nos metadados ou nos arquivos do dataset (adição, remoção) leva a criação de uma nova versão, com identificação global persistente e citação para essas versões [PAC4]. Edições mínimas em metadados levam a novas versões intermediárias sem que a citação seja alterada.

Um fator importante para a preservação digital é representar itens em pacotes que são independentes do *software* e que ficam armazenados em uma estrutura única (contínua), no disco. Dessa forma, a obtenção e a preservação de um item não fica dependente do *software* que constrói e gerencia este item. Dataverse armazena informações da estrutura do pacote (dataset) em base de dados relacional, isto é, armazena pacotes de forma dependente do *software* [PAC5]. Entretanto, Dataverse permite a exportação dos metadados de um dataset (não incluindo os arquivos do datset) no formato DDI Codebook, que resulta em um arquivo XML que descreve todo o pacote, incluindo metadados estruturais (estruturas físicas e lógicas dos documentos, incluindo variáveis em documentos tabulares). Dataverse não exporta o pacote, isto é, seus dados e metadados em uma estrutura única, mas disponibiliza uma API em através da qual todas as informações que compõem o dataset podem ser objtidas. A Figura 20 exemplifica, esquematicamente e resumidamente, a representação em DDI Codebook do dataset com DOI 10.18738/T8/XN2POZ, armazenado no Repositório Texas Data Library [15].

Figura 20 - Metadados em DDI CodeBook



Fonte: Adaptado de TDL - https://dataverse.tdl.org/dataset.xhtml?persistentId=-doi:10.18738/T8/XN2POZ

Dataverse permite a publicação de dados em formatos não proprietários e legíveis por máquina (dados 3 estrelas). Também possibilita o armazenamento de arquivos no formato RDF/XML, isto é, arquivos que contêm dados representados no padrão da Web Semântica (4 estrelas) e interligados (5 estrelas). Não permite o armazenamento, o gerenciamento e a manipulação de triplas RDF, pois não armazena e gerencia triplas RDF (triple store) [PAC6].

4.2.4 Descrição e Documentação dos Conjuntos de Dados

Essa seção analisa Dataverse com relação aos critérios: informação de proveniência e de contextualização da produção dos dados [DOC1], informação descritiva de ações de gestão e preservação digital (metadados administrativos) [DOC2], informação descritiva sobre o conteúdo intelectual (metadados descritivos) [DOC3], informação descritiva sobre a estrutura de representação dos dados (metadados estruturais) [DOC4], informação descritiva sobre aspectos técnicos dos objetos digitais (metadados técnicos) [DOC5], vocabulários controlados [DOC6], recursos para representar novos esquemas de metadados e estender esquemas existentes [DOC7], recursos para realizar mapeamentos (crosswalks) entre esquemas de metadados ou gerenciar metadados representados em múltiplos formatos [DPC8], descrever os conjuntos de dados em ambiente Linked Data/Web Semântica [DOC9], descrever os conjuntos de dados integrados com Sistemas de Informação da Pesquisa [DOC10] e documentação dos dados [DOC11].

Dataverse possui seus esquemas próprios de metadados, mas que são mapeados e exportados para esquemas de metadados que são apropriados para repositórios multidisciplinares e das áreas da Ciências Sociais e Humanidades, Astronomia e Astrofísica e Ciências da Vida [DOC3]. Além desses esquemas fornecidos por Dataverse, a ferramenta também permite a criação de novos elementos de metadados e de vocabulários controlados para valores desses elementos e de outros elementos já existentes.

Para descrever e documentar um conjunto de dados (Figura 22), Dataverse desenvolveu um esquema de metadados, chamado de Citação, de uso obrigatório, adequado para repositórios multidisciplinares, e já seguindo recomendações para a área das Ciências Sociais e Humanidades.

Quadro 4 - Metadados de Citação de Dataverse

• Título, Subtítulo, Título Alternativo, URL alternativa, Outro ID • Autor, Contato, • Produtor, Data de produção, Local de Produção • Colaborador • Distribuidor, Data de distribuição • Depositante, Data de Depósito	• Assunto, Palavra chave • Publicações Relacionadas • Notas • Idioma • Período de tempo coberto • Grant • Descrição • Data de coleção • Tipo de dado • Series	• *Software* • Material relacionado, datasets relacionados, outras referências • Fontes de Dados, Origem das Fontes, Características das Fontes, Documentação e acesso às fontes

Fonte: Dataverse Metadata v4.x - https://docs.google.com/spreadsheets/d/13HP-jI_cwL-DHBetn9UKTREPJ_F4iHdAvhjmlvmYdSSw/edit#gid=0

O Quadro 4 apresenta os principais elementos do esquema de Citação de Dataverse. Esse esquema é compatível e exportável para os padrões DDI Lite[137], DDI 2.5 Codebook [130] e Dublin Core. Dataverse, ao permitir o mapeamento de seus metadados para Dublin Core e DataCite, dá suporte à descoberta de informações. Ao mapear seus metadados para DDI Codebook (como exemplificado na Figura 20), Dataverse dá suporte à representação de um estudo, incluindo as suas variáveis e aos seus arquivos, conforme Ciências Sociais e Humanidades [DOC3].

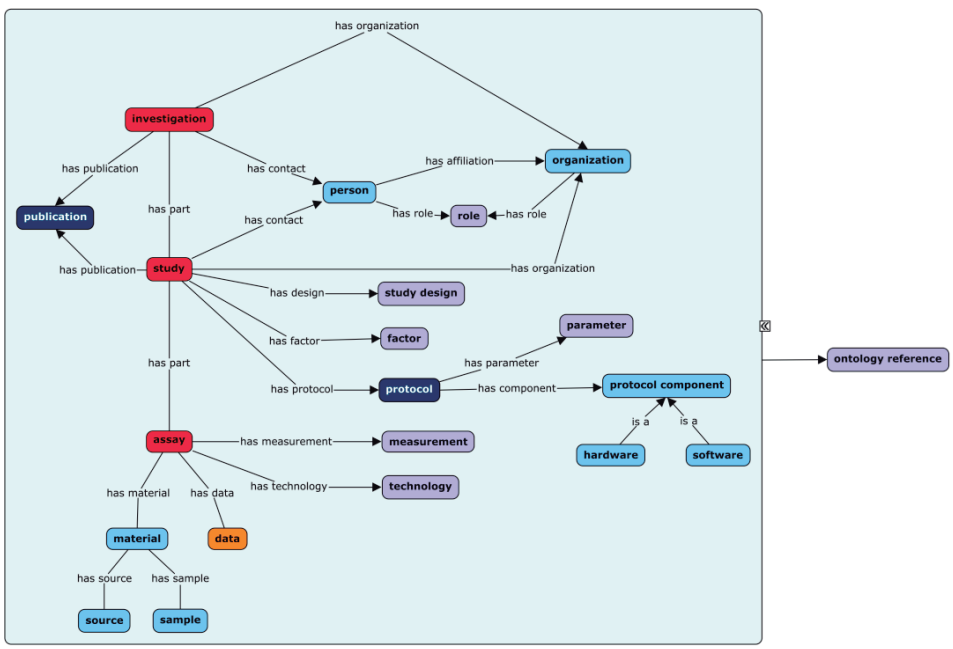
A ferramenta disponibiliza outros esquemas de metadados, que podem ser usados em um dataset, juntamente com o esquema de Citação, que é obrigatório. Os esquemas disponibilizados permitem descrever dados Geoespaciais, da Astrofísica e Astronomia e das Ciências da Vida [DOC3]. O esquema de metadados de Dataverse para Astronomia e Astrofísica permite que os elementos descritos sejam mapeados e exportados para VOResource Schema format[138], esquema desenvolvido para descrever recursos de observações virtuais (Virtual Observatory).

137 DataCite 3.1 - https://schema.datacite.org/meta/kernel-3.1/

138 VOResource Schema - http://www.ivoa.net/documents/VOResource/20180625/index.html

O esquema de Ciências da Vida de Dataverse é baseado na especificação ISA-Tab[139], que foi definido de acordo com ISA-Framework. Esse esquema envolve a descrição de dados experimentais, envolvendo características de amostras, tecnologias e tipos de medição, relações entre amostras e dados. A Figura 21 apresenta o modelo abstrato que é base para ISA-Tab.

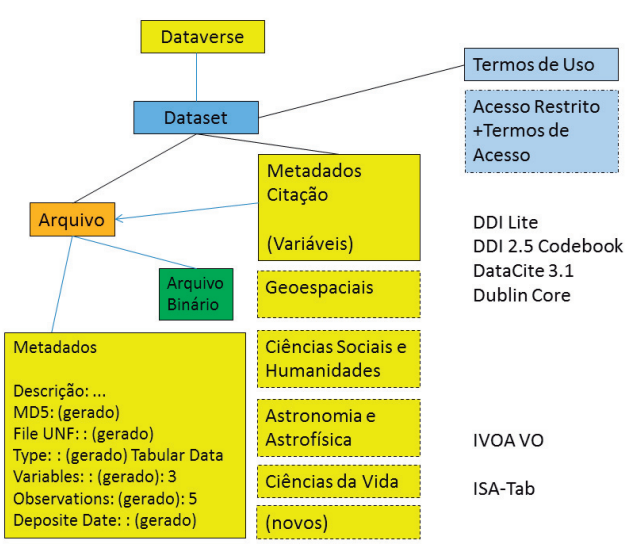
Figura 21 - ISA Abstract Model



Fonte: https://isa-specs.readthedocs.io/en/latest/isamodel.html

Metadados que indicam termos de uso dos dados, como licença, são representados no componente **Termos** de um dataset (Figura 26). Dataverse também disponibiliza metadados para descrever os arquivos (como descrição, formato de arquivo, checksum (MD5, UNF) e as variáveis, em caso de estruturas tabulares) conforme demonstrado na Figura 22.

Figura 22- Metadados de um Dataset



Fonte: Dados da pesquisa

139 ISA-Framework - ISA-Tab - https://www.isacommons.org/

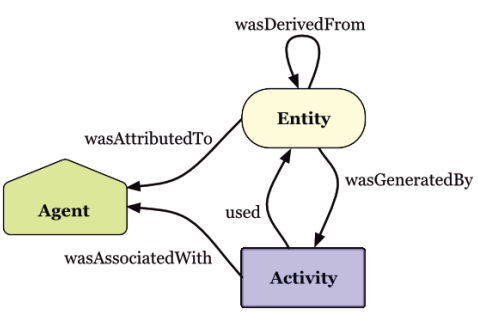
O esquema de metadados de Citação de Dataverse possui elementos de proveniência, que permitem descrever autor, produtor, colaborador, depositante, fontes (Quadro 4), por exemplo. Ao gerenciar versões, o ambiente permite rastrear todas as mudanças que ocorreram no dataset ao longo de seu ciclo de vida, isto é, proveniência [DOC1]. O ambiente registra as versões de um dataset, os publicadores, as data de publicação e as mudanças ocorridas. A Figura 23 apresenta informações de proveniência registradas para as versões do dataset com DOI 10.7910/DVN/28809.

Figura 23 - Proveniência: Registro das Versões de Dataset

Fonte: Harvard Dataverse - https://doi.org/10.7910/DVN/28809

Para arquivos depositados em datasets, Dataverse possibilita a inclusão de informações específicas de proveniência através da descrição textual da proveniência e/ou da inclusão (importação) de um arquivo com metadados de proveniência[140]. Nesse arquivo, os metadados de proveniência devem estar de acordo como o padrão W3C Prov[141]. Esse padrão define um modelo básico para construir representações de entidades, pessoas e processos envolvidos em produzir um componente de dado (Figura 24).

Figura 24 - Estrutura básica de PROV



Fonte: W3C. PROV - https://www.w3.org/TR/prov-primer/

Com relação aos metadados técnicos [DOC 5], Dataverse usa JOHVE [53] para verificar e identificar os formatos dos arquivos, mas não extrai e registra características técnicas desses formatos. Dataverse registra o formato dos arquivos em metadados.

140 Dataverse User Guide - Data Provenance http://guides.dataverse.org/en/latest/user/dataset-management.html
141 W3C Prov - https://www.w3.org/TR/prov-primer/

Com relação a metadados de preservação digital [DOC2], Dataverse registra ações relativas à submissão, como quem depositou, quem publicou e a data do depósito. Dataverse aplica a função *hash* (MD5), armazenando o produto dessa função em metadados (metadado de fixidez). Para dados tabulares, Dataverse adota o Universal Numerical Fingerprint (UNF) [142], 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.

Com relação a metadados estruturais [DOC4], Dataverse representa as informações das estruturas dos datasets em banco de dados relacional. Dataverse permite a exportação de metadados no formato DDI Codebook, que envolve também a descrição dos componentes de um estudo (dataset), incluindo seus arquivos e as variáveis que são representadas nesses arquivos (Figura 20 e Figura 25). No caso de dados tabulares, Dataverse possui micro serviços que extraem informações de arquivos nos formatos csv, excel, Stata, SPSS ou R e geram metadados que descrevem as variáveis presentes nesses arquivos. A Figura 25 apresenta, em DDI Codebook, os metadados que descrevem as variáveis presentes nos arquivos que contêm dados tabulares do dataset com DOI 10.7910/DVN/ZHCUUU.

Figura 25 - Descrição de Variáveis em Dados Tabulares em DDI


```
−<codeBook xsi:schemaLocation="ddi:codebook:2_5 http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd" version="2.5">
  +<docDscr></docDscr>
  +<stdyDscr></stdyDscr>
  +<fileDscr ID="f3135273" URI="https://dataverse.harvard.edu/api/access/datafile/3135273"></fileDscr>
  +<fileDscr ID="f3135270" URI="https://dataverse.harvard.edu/api/access/datafile/3135270"></fileDscr>
  −<dataDscr>
    −<var ID="v18444670" name="BG_ID_00" intrvl="discrete">
        <location fileid="f3135273"/>
        <labl level="variable">BG_ID_00</labl>
        <sumStat type="mode">.</sumStat>
        <sumStat type="stdev">418846.49475432356</sumStat>
        <sumStat type="vald">525.0</sumStat>
        <sumStat type="medn">2.50250817003E11</sumStat>
        <sumStat type="max">2.50251404007E11</sumStat>
        <sumStat type="invd">0.0</sumStat>
        <sumStat type="mean">2.5025073591372 19E11</sumStat>
        <sumStat type="min">2.50250001001E11</sumStat>
        <varFormat type="numeric"/>
        <notes subject="Universal Numeric Fingerprint" level="variable" type="Dataverse:UNF">UNF:6:SoOPA/537dnz4HT2I26qQA==</notes>
      </var>
    +<var ID="v18444672" name="soccoh_0810" intrvl="contin"></var>
    +<var ID="v18444669" name="soccon_0810" intrvl="contin"></var>
    +<var ID="v18444665" name="gangact_0810" intrvl="contin"></var>
    +<var ID="v18444666" name="IntgenClos_0810" intrvl="contin"></var>
    +<var ID="v18444668" name="Abuse_0810" intrvl="contin"></var>
    +<var ID="v18444661" name="NbhdInv_0810" intrvl="contin"></var>
    +<var ID="v18444667" name="PhysDis_0810" intrvl="contin"></var>
    +<var ID="v18444664" name="SocDis_0810" intrvl="contin"></var>
    +<var ID="v18444671" name="Police_0810" intrvl="contin"></var>
    +<var ID="v18444663" name="RecipExch_0810" intrvl="contin"></var>
    +<var ID="v18444662" name="Unsafe_0810" intrvl="contin"></var>
    +<var ID="v18444673" name="SocNet_0810" intrvl="contin"></var>
    +<var ID="v18444660" name="CollEff_0810" intrvl="contin"></var>
    +<var ID="v18444655" name="tweet_id" intrvl="discrete"></var>
    +<var ID="v18444648" name="time" intrvl="discrete"></var>
    +<var ID="v18444650" name="lat" intrvl="contin"></var>
    +<var ID="v18444651" name="lon" intrvl="contin"></var>
    +<var ID="v18444647" name="goog_x" intrvl="contin"></var>
    +<var ID="v18444654" name="goog_y" intrvl="contin"></var>
    +<var ID="v18444659" name="sender_id" intrvl="discrete"></var>
    +<var ID="v18444656" name="sender_nam" intrvl="discrete"></var>
    +<var ID="v18444653" name="source" intrvl="discrete"></var>
    +<var ID="v18444652" name="reply_to_u" intrvl="discrete"></var>
    +<var ID="v18444649" name="reply_to_t" intrvl="discrete"></var>
    +<var ID="v18444657" name="place_id" intrvl="discrete"></var>
    +<var ID="v18444658" name="tweet_text" intrvl="discrete"></var>
  </dataDscr>
  +<otherMat ID="f3135272" URI="https://dataverse.harvard.edu/api/access/datafile/3135272" level="datafile"></otherMat>
</codeBook>
```


Fonte: Harvard dataverse - https://doi.org/10.7910/DVN/ZHCUUU

142 Micah Altman and Gary King. 2007. "A Proposed Standard for the Scholarly Citation of Quantitative Data." D-Lib Magazine, 13. http://www.dlib.org/dlib/march07/altman/03altman.html

Dataverse permite a definição de vocabulários controlados para o preenchimento de valores na produção de metadados. Esses vocabulários são cadastrados através de uma tabela [DOC6]. Dataverse não dispõe de recursos (interface de usuário) para o cadastramento e gerenciamento desses vocabulários.

Dataverse permite a exportação dos metadados de um dataset em diversos esquemas de metadados, como DDI Codebook, DDI Lite, Dublin Core, DataCite. Entretanto não dispõe de recursos que permitam a construção de regras de mapeamento (*crosswalks*) dos metadados armazenados para outros esquemas [DOC8].

Dataverse, na sua distribuição, não dispõe de recursos para a colheita de metadados em documento RDF/XML, padrão da Web Semântica, e não possibilita o armazenamento e manipulação de metadados na forma de triplas RDF [DOC9]. Dataverse possui alguns metadados em comum com Sistemas de Informação de pesquisa, como projeto/grant e produtor (Quadro 4) [DOC10].

Dataverse não dispõe de recursos específicos para gerenciar a documentação que dá apoio ao uso de dados. Entretanto, permite que um arquivo de documentação 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. [DOC11].

4.2.5 Produção dos Conjuntos de Dados

Essa seção analisa Dataverse com relação aos critérios: gerenciamento de acordos de submissão e licença [SUB1], fluxos de submissão [SUB2], transferência legal da custódia dos dados ao repositório [SUB3], validação e verificação pacotes de submissão [SUB4], extração, verificação e/ou produção de metadados técnicos, administrativos e descritivos [SUB5], submissão em lote, por máquina e em ambientes distribuídos [SUB6], uso de padrões [SUB7].

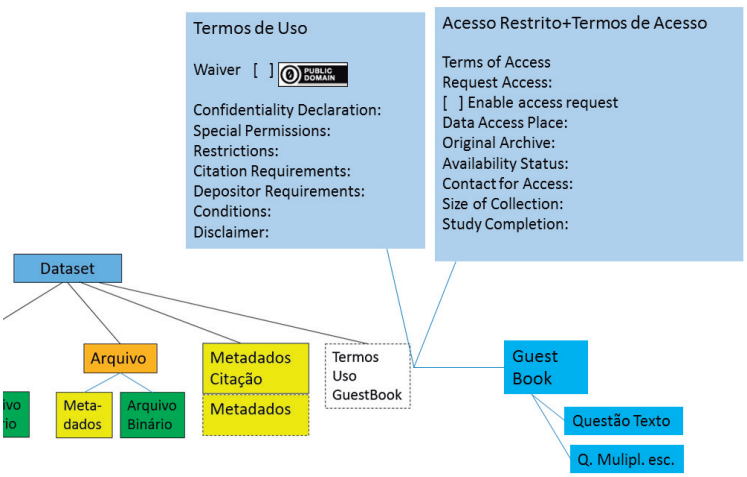
O ambiente não possui funcionalidades de gerenciamento de licenças e contratos (acordos de submissão), isto é, que permitam o cadastramento e a gestão de licenças e acordos, e seu relacionamento com datasets que são regidos por esses documentos [SUB1].

Acordos e licenças são operacionalizadas no dataset por meio do componente Termos (Figura 26) [SUB1]. Em Termos, produtores informam a licença de uso, permissões/restrições especiais, requisitos para citação e do depositante, e isenção de responsabilidade (*disclaimer*). O ambiente dispõe, como padrão, da licença Creative Commons CC0, mas abre a possibilidade para que seja informado um texto descrevendo uma licença alternativa. Somente usuários com perfis autorizados podem definir os termos.

Padrões de licenças podem ser definidos e aplicados em vários dataset por meio do recurso **Template**. Por meio desse recurso, uma configuração padrão (*default*) para dataset é criada, contendo valores pré-definidos para alguns metadados, incluindo termos

(seus metadados e sua licença). Um Template pode ser usado em um ou mais dataverses. Com isso, a criação de datasets desses dataverses irá obedecer às configurações iniciais estabelecidas nos Templates, incluindo informações relativas aos termos [SUB1].

Figura 26 - Termos e Guestbook



Fonte: Dados da pesquisa

Dataverse também permite a definição de termos para acesso aos dados, para casos em que arquivos têm acesso restrito. Para estes arquivos, o ambiente dispõe de metadados que informam de que maneira os usuários consumidores podem ter acesso aos dados, incluindo local para acesso, estado (como embargo), contrato para acesso (Figura 26).

Para submissão e para aprovação (publicação) de um dataset, políticas e estratégias de submissão são construídas por meio da criação e da atribuição de papéis a grupos de usuários. Dataverse permite, por exemplo, que um grupo de usuários exerça a função de criar ou de editar um dataset, e outro grupo, a função de aprovar e publicar um dataset ou uma nova versão de um dataset. Além disso, Dataverse notifica os envolvidos. [SUB2]

A transferência da custódia dos dados ao repositório ocorre por meio dos recursos que a ferramenta dispõe para incluir e para publicar datasets. Produtores são devidamente autenticados, e usuários com papéis devidamente atribuídos têm autorizaçoes para submeter e para publicar conjuntos de dados [SUB3]. O ambiente registra as versões, com data e pessoa que publicou.

A validação e a verificação de um pacote submetido via interface de usuário é realizada por meio dos seguintes mecanismos: Dataverse disponibiliza uma interface de usuário para construção do pacote de submissão, direcionado o usuário a produzir um pacote, isto é, um dataset. O pacote (dataset) assume versão "DRAFT" até que seja publicado por alguém, que pode ser incumbido de tarefas de verificação e de validação [SUB4]. O uso de Templates permite a definição de padrões para datasets.

Na submissão, Dataverse executa micro serviços para gerenciar formatos (verificar formatos via JHOVE) e extrair metadados estruturais de arquivos de dados tabulares e geoespaciais [SUB4]. Com relação a dados tabulares, ao ler um arquivo do tipo csv, excel, Stata, SPSS ou R, o Dataverse analisa a integridade desse arquivo e gera metadados que descrevem as variáveis neles representadas. Além disso, gera uma outra representação para esse arquivo (no formato .tab), com o objetivo de ter uma representação única (canônica) para todos dados tabulares depositados. Com isso, Dataverse permite a integração de dados tabulares com ferramentas de manipulação, como TwoRavens [131].

Dataverse trata (verifica) dados geoespaciais vetoriais (*shapefiles* - usados por Sistemas de Informações Geográficas) e permite sua integração com ambiente WorldMap [132], isto é, dados geoespaciais armazenados em Dataverse podem ser explorados por WoldMap. Ao receber um pacote (arquivo .zip) que contém arquivos que representam um *shapefile* (nos formatos .shp, .shx, .dbf, .prj, entre outros), Dataverse desempacota o arquivo, verifica os arquivos do pacote e gera um novo pacote, seguindo seus padrões de agrupamentos, que será também armazenado. [SUB4].

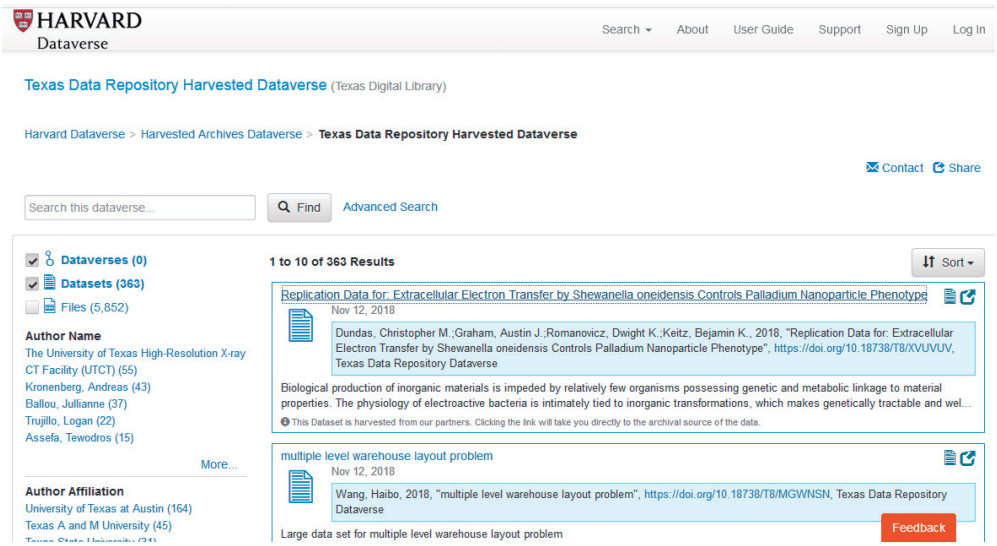
Dataverse permite também que dados tabulares que contêm informações geoespaciais possam ser mapeados para serem explorados pela ferramenta WordMap (Figura 19). Através de uma interface interativa, produtores podem mapear variáveis de dados tabulares (que representam informações geoespaciais) para estruturas que permitam sua análise em WordMap [SUB4].

Dataverse não dispõe na sua distribuição de micro serviços que extraem automaticamente metadados técnicos dos arquivos submetidos (como informações sobre formato) [SUB4] [SUB5].

Dataverse permite a submissão por máquinas via rede por meio do protocolo Sword. Também disponibiliza uma API que permite o gerenciamento (criação, edição, remoção) de dataverses, datasets e arquivos, assim como de usuários, papéis e grupos.

Dataverse também possibilita que coleções sejam alimentadas a partir da colheita de metadados de dados armazenados em outros repositórios, via protocolos OAI-PMH. Isso permite o uso de Dataverse como agregador de informações em um ambiente federado de repositórios [SUB6]. Por exemplo, Harvard Dataverse possui o dataverse Texas Data Repository Harvested Dataverse (Figura 27). Os datasets desse dataverse são colhidos do repositório Texas Data Repository.

Figura 27 - Colheita em Dataverse



Fonte: Harvard dataverse - https://dataverse.harvard.edu/dataverse/tdr_harvested

Dataverse possui uma extensão que permite sua integração como o *software* para editoração e publicação de revistas eletrônicas Open Journal System (OJS)[143]. Essa extensão permite que dados que estão associados à submissão de um artigo sejam automaticamente armazenados em Dataverse, no momento que o artigo for aceito através dos fluxos e operações do ambiente OJS do periódico.

4.2.6 Armazenamento a Longo Prazo e Planejamento da Preservação

Essa seção analisa Dataverse com relação aos critérios: Serviços/microserviços para garantir acesso a longo prazo [PD1], planejamento e ações de preservação digital [PD2], integração com serviços de preservação digital de terceiros ou colaborativos [PD3], exclusão de dados [PD4] e uso em Repositório Digital Confiável certificado [PD5].

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, Dataverse extrai as variáveis, gera uma representação dos dados em formato canônico (.tab), e extrai e armazena o Universal Numeric Fingerprint (UNF), número *hash* criptografado (assinatura), que pode ser usado para identificar unicamente uma versão de conjunto de dados [PD1].

Além de realizar a checagem da integridade dos objetos na submissão, é importante que um *software* de repositório ofereça micro serviços de preservação digital que atuem periodicamente ou, quando necessários, sobre os objetos armazenados, como micro serviços que realizam checagem de integridade, verificação de ocorrência de vírus, migrações de formatos, controle de mídias etc. Dataverse não dispõe desses serviços [PD1].

143 Dataverse OJS - https://projects.iq.harvard.edu/ojs-dvn/book/project-documentation

Com relação ao apoio à preservação digital, Dataverse não disponibiliza de funcionalidades que permitem a definição e gestão de políticas de formatos (com a indicação e o controle, por exemplo, sv formatos que serão permitidos e daqueles que que não serão aceitos) [PD2].

Na distribuição atual de Dataverse não existem referências[144] sobre participação de Dataverse em redes cooperadas de preservação digital que usam o protocolo Lockss [57] [PD3]. Entretando, Lockss é apresentado na literatura e em guias de versões anteriores como característica de Dataverse [PD3]. [145] [126]

Uma estratégia alternativa para realizar a preservação digital dos objetos armazenados é exportar os objetos e os metadados, em pacotes, para ambientes de arquivamento, responsáveis pela preservação digital. Em Dataverse, essa alternativa está em discussão[146]. Dataverse disponibiliza de uma rotina (script) para realizar cópias de segurança.

Como alternativa, Dataverse pode ser integrado com Arquivematica[147] [PD3], *software* que realiza serviços e planejamento da preservação digital, através de módulo de integração com Dataverse de Archivematica (Figura 28). Nesse caso, Archivematica assume a responsabilidade de preservar datasets armazenados em Dataverse, ficando com Dataverse a somente função de dar acesso aos dados. Para preservar um dataset a longo prazo, através do módulo de transferência de Archivematica, uma cópia do dataset é transferida de Dataverse para Archivematica. Archivematica usa a API de Dataverse para buscar o dataset (suas estruturas, metadados e arquivos) e gera um Pacote de Submissão de Informação (SIP), que segue as estruturas recomentadas de preservação digital adotadas por Archivematica. Esse pacote é então ingerido por Archivematica, sendo transformado em Pacote de Armazenamento (AIP), via micro serviços [PD1]. O pacote AIP é armazenado e gerenciado por Archivematica visando sua preservação a longo prazo [PD2].

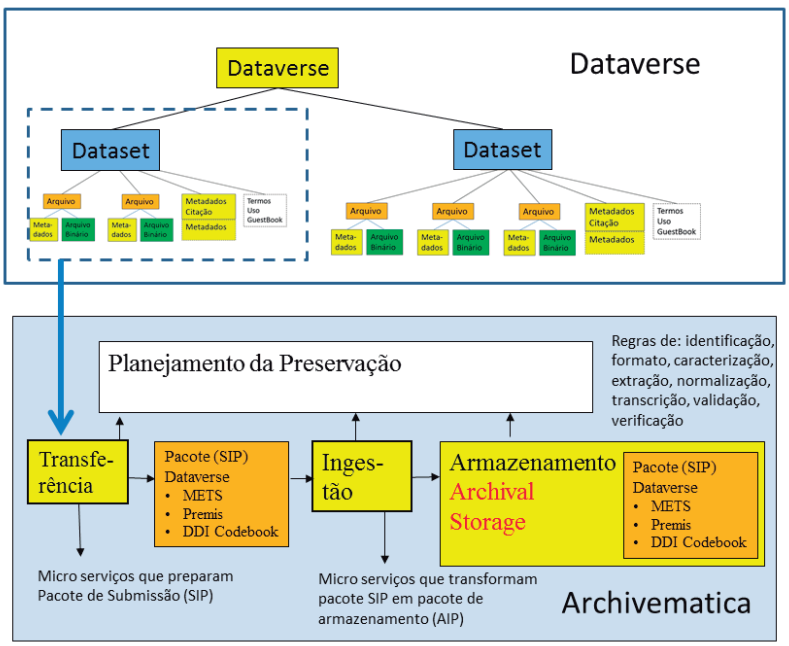
144 Dataverse – Guides - http://guides.dataverse.org/en/latest/search.html?q=lockss

145 The Harvard Dataverse Network 3.6 documentation – User Guide - http://dvn.library.ubc.ca/guides/dataverse-user-main.html

146 Data and Metadata Packaging for Archiving. https://github.com/QualitativeDataRepository/dataverse/wiki/Data-and-Metadata-Packaging-for-Archiving

147 Archivematica – Dataverse - https://wiki.archivematica.org/Dataverse

Figura 28 - Integração Dataverse e Archivematica



Fonte: Dados da pesquisa

Dataverse permite que dados sejam embargados e removidos [PD4]. Datasets removidos não são excluídos fisicamente. O embargo é gerenciado através 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. Dataverse não gerencia período de embargo.

Repositórios digitais que usam Dataverse obtiveram certificações Data Seal Approval [148] e Core Trust Seal[149], como Odum Institute Archive Dataverse[150], demonstrando que a ferramenta não traz impedimentos para tal [PD5].

4.2.7 Acesso e Uso dos Conjuntos de Dados

Essa seção analisa Dataverse com relação aos critérios: Recuperação de informação [AC1], informações sobre direitos de uso, licenças [AC2], informações de proveniência e para uso dos dados [AC3], informações de citação [AC4], identificadores globais e persistentes [AC5], identificadores globais e persistentes para recursos relacionados [AC6], acesso aos dados e aos metadados via identificador e protocolo aberto [AC7], restrições de acesso [AC8], Gerenciamento e autenticação de usuários envolvidos com acesso [AC9], entrega de dados ao consumidor [AC10], entrega dos metadados ao con-

148 Repositórios com Certificação DSA, cadastrados no Diretório Re3data - https://www.re3data.org/search?query=&software%5B%5D=DSpace&certificates%5B%5D=DSA

149 Repositórios com Certificação Core Trust Seal, cadastrados no Diretório Re3data - https://www.re3data.org/search?query=&software%5B%5D=DSpace&certificates%5B%5D=CoreTrustSeal

150 Repositório ODUM - https://dataverse.unc.edu/dataverse/odum

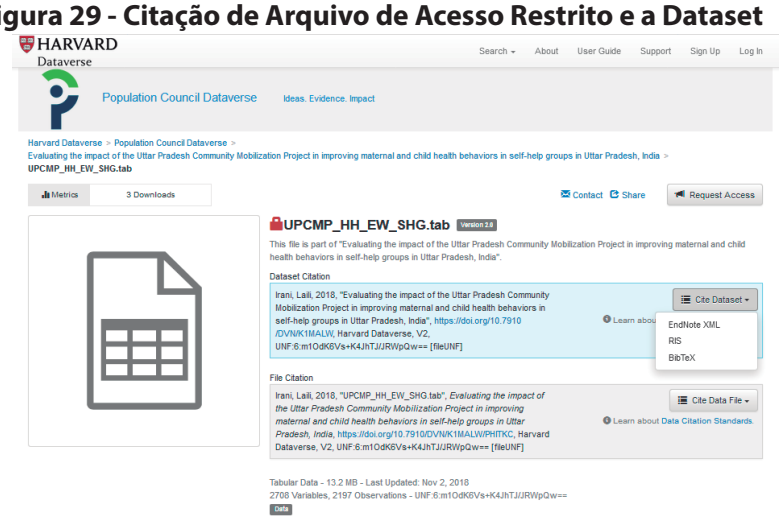
sumidor [Ac11], ferramentas para visualização e análise de dados [AC12], estatísticas e relatórios de uso [AC13], acesso às descrições em formatos para Linked Data/Web Semântica [AC14], recuperação em ambiente Linked Data/Web Semântica [AC15].

Dataverse permite busca básica e avançada, e navegação por facetas. A busca básica recupera informação nos conteúdos, incluindo dataverses, datasets e arquivos. A busca avançada permite a recuperação por campos, de dataverses, de datasets e de arquivos. Em arquivos com dados tabulares, Dataverse permite a recuperação por variáveis (descritas através dos metadados). Os resultados das buscas podem ser ordenados por relevância, nome e 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 dataverse pode ser configurado para exibir um conjunto específico facetas [AC1].

Através do componente Termos, de um dataset, Dataverse disponibiliza aos usuários as licenças de uso e as informações de como ter acesso a dados que são de acesso restrito [AC2] (Figura 26). Dataverse disponibiliza informações de proveniência presentes nos metadados de citação (como autor, depositante, data). Também apresenta informações de proveniência relevante às versões do dataset (data e publicador), assim como metadados de proveniência associados a arquivos (Figura 23) [AC3].

Dataverse dá um grande destaque à citação [AC4]. Os metadados obrigatórios são compatíveis com o padrão de DataCite, permitindo a integração com serviços de citação, como DataCite, e também a geração de informação de citação a ser disponibilizada ao usuário. O padrão de citação usado em Dataverse oferece tanto o reconhecimento adequado dos autores, como o uso de identificadores globais persistentes. Usa impressões digitais (UNF) para permitir a verificação da autenticidade dos dados citados. A Figura 29 exemplifica a citação de um arquivo e de seu de dataset. Nessas citações, observamos a informação da versão e o uso de elementos como DOI e UNF. O ambiente também permite a exportação da citação em formatos EndNote, RIS e BibTex.

Figura 29 - Citação de Arquivo de Acesso Restrito e a Dataset

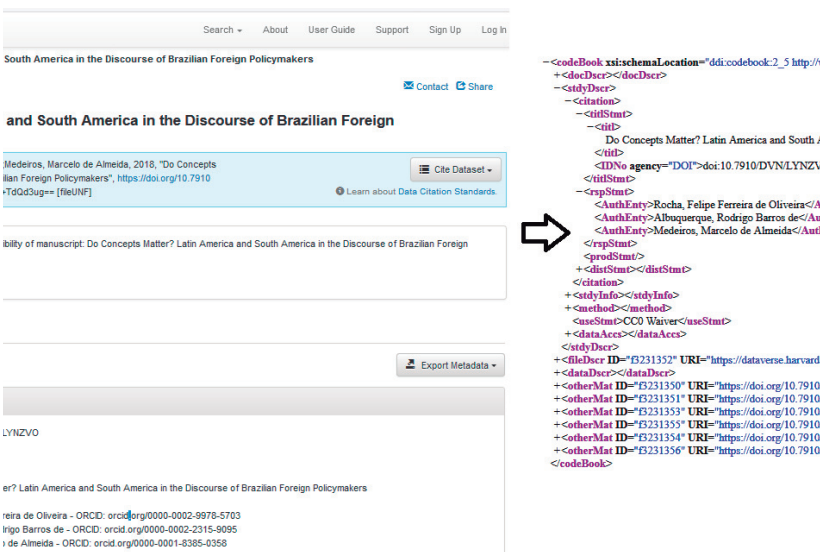


Fonte: Harvard Dataverse - https://doi.org/10.7910/DVN/K1MALW/PHITKC

Em Dataverse, objetos e metadados são acessados via serviços de identificadores globais e via protocolos abertos: permite o uso dos serviços DOI e HandleSystem para identificação, e HTTP e HTTPS, para acesso e transferência de informações entre o cliente e o repositório. [AC5] [AC7]. O uso de DOI implica em registrar cada identificador DOI em uma agência de registro DOI, como DataCite[151] e EZID[152]. Permite que DOIs sejam registrados em DataCite ou EZID, através de funções que chamam as APIs de registros dessas agências.

Dataverse possibilita que o identificador ORCID seja representado junto aos nomes de pessoas, como depositantes e autores de datasets. Entretanto, não observa o uso do serviço de ORCID no apoio ao controle de autoridade [AC6]. A Figura 30 exemplifica um dataset em que seus autores são apresentados juntamente com suas identificações ORCID.

Figura 30 – ORCID em Autores de Dataset



Fonte: Harvard Dataverse https://doi.org/10.7910/DVN/LYNZVO

Dataverse permite acesso aos dados e aos metadados via identificador (Handle ou DOI) e protocolo aberto (Http, Https) [AC6]. Datasets e dataverses, quando publicados, são de acesso público. Arquivos publicados podem ter acesso restrito, como no exemplificado na Figura 29. Nesse caso, o acesso só é possível por usuários autenticados e autorizados [AC8].

Dataverse dispõe do recurso "Livro de Visitas" (Guestbook) para estabelecer um relacionamento com os usuários dos dados. Através desse recurso (Figura 26), 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 serem usados como Livros de Vista de vários datasets, contendo questões de texto livre e múltipla escolha.

Os usuários podem ser autenticados pelo próprio do Dataverse e também pelo uso dos protocolos Shibboleth e OAuth2. Através de OAuth2 é possível usar Google,

151 DataCite – Atribuindo DOI - https://www.datacite.org/dois.html

152 EZID – DOI - https://ezid.cdlib.org/

Facebook e ORCID para autenticar usuários. Shibboleth é usado para *login* institucional. Dataverse NL é um exemplo de repositório que usa autenticação via Shibboleth, provido pelo serviço SURFconext[153] [AC9].

Em Dataverse, documentos e conjuntos de dados podem ser representados em diferentes formatos, a fim de atender demandas de variados consumidores [AC10]. O ambiente não atende ao protocolo OAI-ORE, que permite a obtenção dos datasets por sistemas informatizados e distribuídos. Como alternativa, sistemas informatizados podem obter informações de dataverse através de uma API nativa[154] [AC10]. Já metadados podem ser colhidos através do protocolo OAI-PMH [AC11] e exportados. A Figura 31 apresenta os esquemas em que metadados podem ser colhidos em Harvard Dataverse: Dublin Core, DDC Codebook 2.5.

Figura 31 - Esquemas de Medadados para Colheita em Harvard Dataverse

```
−<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2018-11-14T01:14:20Z</responseDate>
  <request verb="ListMetadataFormats">https://dataverse.harvard.edu/oai</request>
 −<ListMetadataFormats>
  −<metadataFormat>
     <metadataPrefix>oai_dc</metadataPrefix>
     <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>
     <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
   </metadataFormat>
  −<metadataFormat>
     <metadataPrefix>oai_ddi</metadataPrefix>
    −<schema>
       http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd
     </schema>
     <metadataNamespace>ddi:codebook:2_5</metadataNamespace>
   </metadataFormat>
  −<metadataFormat>
     <metadataPrefix>dataverse_json</metadataPrefix>
     <schema>JSON schema pending</schema>
    −<metadataNamespace>
       Custom Dataverse metadata in JSON format (Dataverse4 to Dataverse4 harvesting only)
     </metadataNamespace>
   </metadataFormat>
 </ListMetadataFormats>
</OAI-PMH>
```

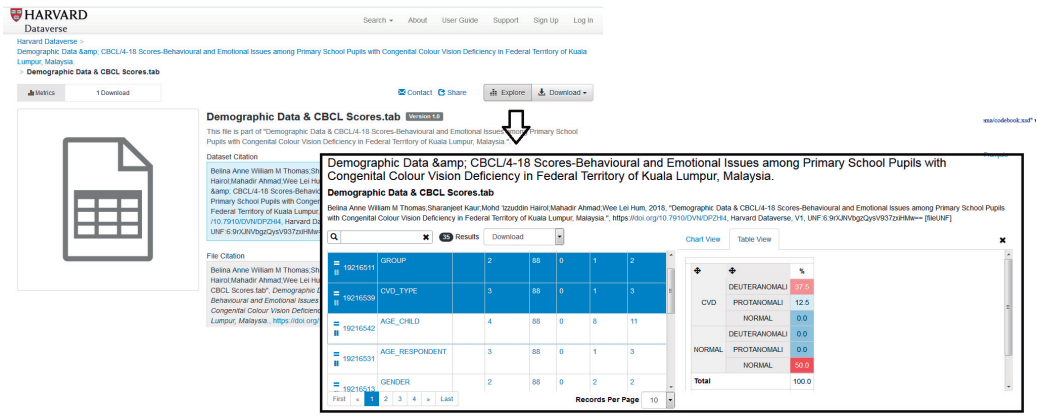

Fonte: Repositório Harvard Dataverse https://dataverse.harvard.edu/oai?verb=ListMetadataFormats

Dataverse oferece recursos para integração com as ferramentas de análise de dados WordMap, para exploração de dados espaciais, e TwoRavens, para exploração de dados tabulares. A Figura 32 exemplifica a exploração de dados geoespaciais vetoriais e tabulares via WorldMap. A Figura 32 exemplifica o uso de TwoRavens para explorar dados tabulares armazenados em Harvard Dataverse. Nesse exemplo, são relacionadas as variáveis GROUP e CVD_TYPE. [AC12]

153 SURF Conext - https://blog.surf.nl/en/the-dutch-dataverse-network-an-inter-institutional-collaboration-effort-facilitated-by-surfconext/

154 API Nativa de Dataverse - http://guides.dataverse.org/en/latest/api/native-api.html

Figura 32 – TwoRavens em Dataverse



Fonte:Harvard Dataverse - https://doi.org/10.7910/DVN/DPZHI4/PSOEQR

Dataverse apresenta o número de downloads para cada dataverse, dataset ou arquivo. Disponibiliza também a API de Métricas (dataverses, datasets e arquivos adicionados por mês, dowloads, dataverses por categorias, datasets por assunto). Através dessa API, é possível construir *scripts* que rodam em navegadores e que exibem métricas, como no caso do Repositório Harvard Dataverse[155] [AC13]

Dataverse não disponibiliza colheita de metadados em RDF/XML, padrão da Web Semântica [AC14], e não possibilita o armazenamento e manipulação de metadados na forma de triplas RDF [AC15]. Também mão apresenta integração com Sistemas de Informação da Pesquisa.

4.3 Comparação entre os *software*

Atualmente vários *software* livres são usados para repositórios de dados da pesquisa. São *software* que foram desenvolvidos especificamente para dados de pesquisa (como Dataverse) ou *software* que originalmente foram desenvolvidos para outros propósitos, como para repositório institucional (como DSpace) e para prover a abertura de dados governamentais (como CKAN). Tendo como referência o diretório de dados da pesquisa Res3Data.org, Dspace, Dataverse e CKAN são os *software* mais usados.

Estes *software* 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 dessa solução integrada é que tudo está presente em um único *software*. Traz facilidades para instalação e operação (integrada), e proporciona robustez do ambiente. A desvantagem está na dificul-

155 Metricas de Harvard Dataverse - https://dataverse.org/metrics

dade 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.

Outra solução que vem sendo adotada por alguns repositórios é implementar componentes funcionais de OAIS na forma de serviços independentes, desenvolvidos a partir do reuso de diversos tipos de *software*, que interagem entre si para atender às funções do repositório. Esse tipo de solução oferece maior flexibilidade de adaptação às necessidades do repositório, principalmente quando estes necessitam armazenar dados com características especiais ou que são coletados de forma automática.

EUDAT é um exemplo solução com essas características, que apresenta uma infraestrutura com serviços para replicar dados (B2Safe), para computar dados (B2Stage), para localizar dados (B2Find), para armazenar, compartilhar e publicar dados (B2Share), para definir identificadores globais e persistentes (B2Handle) e para sincronizar e trocar dados (B2Drop).

EUDAT reusa vários *software* na implementação desses serviços, como por exemplo, iRods , que é usado para o armazenamento a longo prazo (em B2Safe), Ivenio, usado para armazenamento (em B2Safe), usado CKAN para busca (em B2Find), Handle System, para identificação (em BD2HJandle), entre outros .

Repositórios de cauda longa normalmente demandam por requisitos mais genéricos, a fim de atender a diversidade da cauda longa dos dados, sendo mais fácil o uso de *software* integrados, como DSpace, CKAN e Dataverse. Já em casos de repositórios voltados a áreas específicas, que possuem particularidades para metadados, representações de dados, procedimentos de coleta, produção e processamento, pode ser mais conveniente a adoção a estratégia de implementar componentes OAIS como serviços.

Dataverse é um *software* integrado para publicação, compartilhamento e armazenamento de dados. 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 que o uso dos dados somente pode ser feito mediante solicitação. Também permite a configuração e uso de esquemas de metadados (compatíveis com DDI Lite, DDI Codebook, DUblin Core, DataCite, VORResource, ISA-Tab), gerencia versões de conjuntos de dados, identifica unicamente 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, submissão por máquinas (Sword) e colheita de metadados (OAI-PMH). É usado por instituições (heiData[156]/Univ.

156 heiDATA. https://heidata.uni-heidelberg.de/

Heidelberg), por grupos de instituições (DataverseNL/Holanda, Abacus[157]/Canada, Texas Digital Library/EUA), e para repositórios temáticos (ICRIAT[158]/Agricultura) e multidisciplinares (Australian Data Archive[159]).

Alguns repositórios que usam Dataverse são repositórios digitais confiáveis certificados, como Australian Data Archive e TiU Dataverse[160]/DatavereNL, demonstrando que o *software* atende a necessidades desse tipo. Dataverse atende a princípios FAIR conforme [161] Wilkinson e outros.

DSpace é um *software* desenvolvido para repositório institucional. Assim como Dataverse, permite a representação de unidades e subunidades autônomas, com a configuração de fluxos específicos de submissão, uso de diversos esquemas de metadados, identificação universal e persistente através de Handle System, cópias distribuídas de segurança (Lockss), submissão por máquina (Sword) e colheita de metadados (OAI-PMH). É usado por reconhecidos repositórios de dados, como o repositório multidisciplinar Dryad[162] e o repositorio institucional DataShare[163]. DataShare[164] é um exemplo de repositório confiável, certificado, que usa DSpace, demonstrando que repositórios em DSpace podem ser certificados como confiáveis.

Como Dataverse foi desenvolvido para repositório de dados, a representação e a gestão automatizada dos conjuntos de dados é estruturada através do conceito dataset, que inclui dados, metadados de citação, metadados específicos, documentação adicional, citação, gerenciamento de versões, etc. Já DSpace está estruturado no conceito de coleção de itens, com cada item sendo compostos por pastas (bundles) que contém arquivos (bitstreams). No caso do uso de DSpace para gerenciar dados, é necessário configurar metadados, fluxos e interfaces de usuário para conduzir a submissão de dados, como feito no repositório DataShare. DSpace não gerencia versões, incluindo identificação e citação de conjuntos versionados, como Dataverse.

CKAN é um *software* desenvolvido para abertura de dados governamentais. Nele, conjuntos de dados pertencem a organizações e podem ser organizados em grupos. CKAN permite representar ambientes em que organizações possuem e gerenciam conjuntos de dados, na qual usuários, através dos papéis de administrador, editor e membro, são autorizados a acessar, tornar público, editar e remover conjuntos de dados, conforme apresentado na Figura 33. Conjuntos de dados podem ser organizados em grupos, que servem para agrupar conjuntos de dados que pertencem a um mesmo tema, por exemplo.

157 Abacus Dataverse Network https://abacus.library.ubc.ca

158 Int. Crops Research Institute for the Semi-Arid Tropics- http://dataverse.icrisat.org/

159 Australian Data Archive. https://dataverse.ada.edu.au/

160 TiU - CoreTrustSeal https://www.coretrustseal.org/wp-content/uploads/2018/04/Tilburg-University-Dataverse.pdf

161 Wilkinson, M. D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, vol. 3 (2016). https://www.nature.com/articles/sdata201618

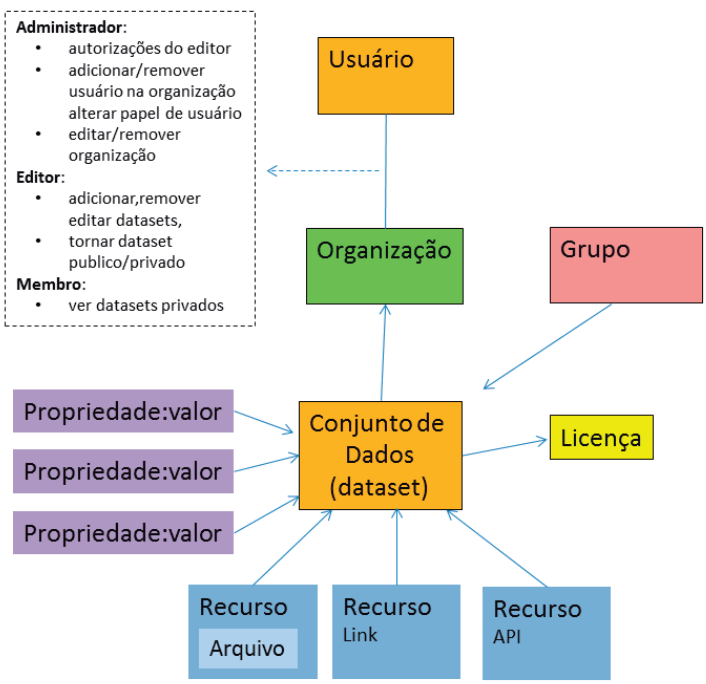
162 Dryad - https://datadryad.org/

163 Edinburgh DataShare. https://datashare.is.ed.ac.uk/

164 Datashare - Data Seal Approval - https://assessment.datasealofapproval.org/assessment_175/seal/pdf/

CKAN possui um pequeno conjunto próprio de metadados, e permite a inclusão de campos (estruturas propriedade:valor) para descrever um conjunto de dados. Possibilita também a seleção de licenças, controla versões de conjuntos de dados, mantém histórico das modificações e gerencia formatos.

Figura 33 - Organização, Grupo, Conjunto de Dados e Papéis em CNAN



Fonte: Dados da pesquisa

Não possui identificador global persistente, nem permite o uso de esquemas de metadados. Mas oferece grande flexibilidade para a criação de extensões, já havendo extensão para identificador persistente (DOI) e para metadados estruturados. Em repositório de dados, o conceito organização de CKAN pode ser usado para representar instituições, unidades ou grupos. Conjuntos de dados podem conter vários arquivos, representando dados e informações adicionais.

CKAN é mais limitado que Dataverse e DSpace no que diz respeito a recursos para organizar unidades e seus dados, para definir políticas de gestão e submissão de dados específicas para cada unidade, assim como no que diz respeito a metadados. CKAN é uma boa solução quando usado como ambiente de publicação e acesso a dados, em que a submissão é feita por outro sistema. É o que ocorre no repositório de dados da Universidade de Bristol[165], no qual pesquisadores solicitam espaço de armazenamento e submetem os dados através do sistema de gerenciamento de pesquisa da universidade. Esse sistema conduz a avaliação dos dados e os publica via CKAN.

165 Repositório Data Bris - https://data.bris.ac.uk/data/

Considerando os três *software*, 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 em DSpace, entretanto exige adaptações configurações, com algumas limitações no controle de versões. CKAN já é mais limitado, entretanto é uma boa alternativa quando usado como serviço de publicação e acesso, com a submissão e preservação digital sendo realizados por outros ambientes.

O Quadro 6 a seguir apresenta comparativamente Dataverse e Dspace com relação aos critérios estabelecidos. O Quadro 5 apresenta os indicadores de atendimento aos critérios apresentados no Quadro 6.

Quadro 5 - Indicadores

Indicador	Cor
Plenamente ou Suficiente	
Suficiente com restrições	
Parcialmente ou muito parcialmente	
Não atende ou Insuficiente	

Quadro 6 - Comparação

Requisito		DSpace	Dataverse
AMB1	Representação do ambiente no qual os conjuntos de dados estão contidos (ambiente organizado em coleções, estudos, grupos, unidades, subunidades, etc.)	■ Grupo/Unidade/ Submunidade: Comunidade ■ Conjuntos de Dados: Coleção ■ Estudo: Item ■ Arquivo: bitstream	■ Grupo/Unidade/Submunidade: Dataverse ■ Estudo: Dataset ■ Arquivo: file
AMB2	Recurso para operacionalizar políticas de funcionamento do ambiente (responsabilidades, grupos de usuários, autenticações, papéis, autorizações para gerenciar e para realizar funções operacionais do repositório)	Recursos para a operacionalização de políticas de funcionamento do ambiente: □ Grupos □ Autorizações à grupos, que permitem a definição de autorizações, criação, remoção, edição e operação em comunidades e coleções	Recursos para a operacionalização de políticas de funcionamento do ambiente: □ Grupos □ Papéis, que permitem a definição de autorizações, criação, remoção, edição e operação em datasets e dataverses □ Autorizações que permitem que grupos exerçam papéis
AMB3	Recursos para estabelecer políticas descentralizadas, em que unidades, grupos ou estudos têm autonomia de gestão e operação	Dispõe de recursos para estabelecer políticas descentralizadas, em nível de comunidade/coleção: □ Politicas atribuídas a comunidades e coleções □ Grupos somente podem ser criados por usuários do grupo administradores	Um dataverse pode ser configurado como um Repositório com gestão independente. Dispõe de recursos para estabelecer políticas descentralizadas, em nível de dataverse/dataset: □ Políticas atribuídas a dataverses e datasets (autorizações, criação, remoção, edicção e operação) □ Grupos podem ser criados por gestores de dataverses

AMB4	Representação de ambiente integrada com Sistemas de Informação de Pesquisa	□ Extensão DSpace-CRIS	□ Não
AMB5	Representação de ambiente para Web Semântica / Dados Abertos e Ligados	□ Módulo em que metadados são replicados em triplestore (base de dados de triplas RDF)	□ Não
AMB6	Recursos que permitam transparência e feedback aos envolvidos, com operações sendo realizadas através de fluxos de trabalhos definidos, permitindo rastreabilidade e auditoria, e com mecanismos de comunicação para manter os envolvidos atualizados. Relatórios e estatísticas.	Permite a definição de fluxo de submissão de itens (que seriam estudos), com as etapas: produção dos metadados e upload dos arquivos, avaliação e conferência do pacote, que são realizadas por pessoas devidamente autorizadas. Registra e dá feedback aos envolvidos. Registra em metadados de proveniência.	Processo de submissão construído a partir da criação de grupos e atribuição de papéis Datasets são criados e submetidos a revisão. Quando aprovados torna-se novas versões Os envolvidos são notificados, metadados registram depositante e produtor.
PAC1	Natureza dos conjuntos de dados (Texto, Multimídia , Modelo/Animações, Simulação, *Software* , Específico de Disciplina, Específico de Instrumento etc.)	Arquivo de qualquer formado: Texto, Tabelas, Multimídia, RDF/XML etc.	□ Arquivo de qualquer formado: Texto, Tabelas, Multimídia, RDF etc.
PAC2	Estruturas para representar os conjuntos de dados em pacotes (pastas, hierarquias de pastas, nomes para pastas e arquivos etc.) compatíveis com estruturas para representar conjuntos de dados de pesquisa	Permite representar conjunto de dados da seguinte forma: □ Conjunto de Dados: Item □ Arquivo: bitstream (em pastas buldle)	Permite representar conjunto de dados da seguinte forma: □ Conjunto de Dados: dataset □ Arquivo: file
PAC3	Formatos de arquivos aceitos (gerenciados pela solução tecnológica)	Aceita variados formatos. Cadastro de formatos, com informação de política (aceito, aceito mas sem responsabilidade com preservação e não aceito). Não verifica formatos, apenas os identifica pela terminação do nome do arquivo.	Aceita variados formatos Não possui cadastro de formatos para definição de políticas. Usa JHOVE para verificar formatos
PAC4	Recursos para versionamento de conjuntos de dados	Não possui versionamento na versão padrão, mas dispõe de uma extensão para versionamento	Permite o versionamento de dataverses, com identificação global e persistente

PAC5	Uso de Padrões (para pacotes, metadados, formatos de arquivo)	Estruturas próprias e dependentes do *software* para representar e descrever pacotes (estrutura do pacote armazenada em base de dados) Exporta pacotes em padrões recomendados para preservação digital (METS+PREMIS	Estruturas próprias e dependentes do *software* para representar e descrever pacotes (estrutura do pacote armazenada em base de dados) Exporta metadados no padrão DDI, que possui a descrição das estruturas dos conjuntos de dados. Metadados e dados do pacote podem ser obtidos via API
PAC6	Qualidade dos dados. Dados 5 Estrelas.	Replica metadados para bases de dados de triplas (RDF). Permite o armazenamento de dados no formato RDF/XML, que podem ser feitos de acordo com os requisitos "5 Estrelas"	Permite o armazenamento de dados no formato RDF/XML, que podem ser feitos de acordo com os requisitos "5 Estrelas"
DOC1	Informação de proveniência e de contextualização da produção dos dados, incluindo versionamento, produtores, processos de produção, recursos relacionados (como publicações), compreensível por máquina e humanos e aceita pela comunidade (metadados administrativos)	As ações realizadas na execução da submissão são armazenadas como metadados de proveniência do item submetido.	Metadados de proveniência no esquema de Citação Informações de proveniência referentes ao versionamento de dataverses Permite a inclusão de informações de proveniência para cada arquivo, na forma textual ou em metadados especificados em PROV (formato JSON)
DOC2	Informação descritiva de ações de gestão e preservação digital, incluindo verificação da integridade do material, compreensível por máquina e humanos e aceita pela comunidade (metadados administrativos)	As ações realizadas na execução da submissão são armazenadas como metadados de proveniência do item submetido.	Registra informações de submissão na base de dados do ambiente

DOC3	Informação descritiva sobre o conteúdo intelectual, compreensível por máquina e humanos e aceita pela comunidade (metadados descritivos)	Disponibiliza Dublin Core Qualificado e Dublin Core Terms, para descrever item (dataset) Permite a construção de esquemas não hierárquicos compatíveis com padrões de dados, como DataCite. Permite a construção de regras de mapeamento (via linguagem XSLT) que permitem o mapeamento dos metadados não hierárquicos em representações em formatos padrões e hierárquicos, como XSLT	Disponibiliza esquema obrigatório chamado de Citação, compatível com os padrões DDI Lite, DDI 2.5 Codebook e Dublin Core. Ao ser compatível com DDI, atende as Ciências Sociais e Humanidades. Disponibiliza outros esquemas adicionais para dados de Astronomia e Astrofísica e Ciências da Vida, compatíveis, respectivamente, com os padrões VO Resource Schema format e ISA-Tab Disponibiliza esquema com metadados geográficos compatíveis com ferramentas de visualização
DOC4	Informação descritiva sobre o conteúdo intelectual, compreensível por máquina e humanos e aceita pela comunidade (metadados descritivos)	Representa as informações das estruturas dos datasets em banco de dados relacional. Permite a exportação de pacotes em formato compatível com recomendações de pacotes para preservação digital, incluindo informação descritiva em METS	Dataverse representa as informações das estruturas dos datasets em banco de dados relacional. Permite a exportação de metadados no formato DDI, que envolve também a descrição dos componentes de um estudo (dataset), incluindo seus arquivos e as variáveis que são representadas nesses arquivos. Em caso de dados tabulares, extrai metadados que descrevem as variáveis e exporta esses metadados no formato DDI
DOC5	Informação descritiva sobre aspectos técnicos dos objetos digitais (formatos de arquivos, versões dos formatos) , compreensível por máquina e humanos e aceita pela comunidade (metadados técnicos)	Descreve formato de arquivos, que é identificado a partir da terminação presente no nome do arquivo Não extrai metadados técnicos	Descreve formato dos arquivos, que é identificado a partir da ferramenta JOHVE Não extrai metadados técnicos
DOC6	Vocabulários controlados (listas de termos, classificações, tesauros)	Permite a definição de listas contendo vocabulários controlados, não dispondo de recursos para o cadastramento e gerenciamento desses vocabulários Permite o uso de SOLR, externo ao DSPace, para controlar vocabulários	Permite a definição de vocabulários controlados, cadastrados através de tabela, não dispondo de recursos para o cadastramento e gerenciamento desses vocabulários.

DOC7	Recursos para definir e estender esquemas de metadados	Permite a criação de esquemas de metadados com elementos no formato: propriedade-valor. Não permite metadados hierárquicos (valores com estruturas complexas)	Permite a criação de esquemas de metadados com elementos no formato: propriedade-valor. Não permite metadados hierárquicos
DOC8	Recursos para realizar mapeamentos (crosswalks) entre formatos ou gerenciar metadados representados em múltiplos formatos	Uso de regras XSLT para mapear os metadados representados no DSpace para outros esquemas Dispõe de regras já prontas para mapeamento de Dublin Core Qualificado para MARC, RDF,...	Dataverse permite a exportação dos metadados de um dataset em diversos esquemas de metadados, como DDI Codebook, DDI Lite, Dublin Core, Datacite. Não apresenta de recursos que permitam a construção de regras de mapeamentos dos metadados armazenados para outros esquemas.
DOC9	Descrever os conjuntos de dados em ambiente Linked Data/ Web Semântica	Permite a colheita de metadados em documento RDF/ XML. Não permita o armazenamento e manipulação de metadados na forma de triplas RD, mas permite que metadados sejam replicados em triplestore	Não
DOC10	Descrever os conjuntos de dados integrados com Sistemas de Informação da Pesquisa	Através da extensão DSpace CRIS	Dataverse possui alguns metadados em comum com Sistemas de Informação de pesquisa, como criador e projeto
DC11	Documentação dos dados	Dados documentados em arquivos armazenados junto ao item	Dados documentados em arquivos armazenados junto ao item Para dados tabulares, Codebook é gerado a partir da extração das variáveis que estão representadas nos arquivos. Codebook é representado em metadados DDC Codebook
SUB1	Gerenciamento de Acordos de Submissão e Licença (definição de acordos de submissão e licenças, gerenciamento de acordos e licenças acordadas).	Não Gerencia Acordos de Submissão e Licenças. Ao criar uma coleção, são cadastrados os textos referentes ao acordo de submissão e à licença. Na submissão, após o produtor aceitar o acordo de submissão, um documento é criado contendo o acordo e armazenado junto ao item	Não Gerencia Acordos de Submissão e Licenças. Disponibiliza metadados de Termos, incluindo licença padrão e outros campos

SUB2	Fluxos de submissão: produção, aprovação de pacotes de submissão. Publicação de pacotes submetidos e aceitos	Fluxo de submissão que pode ser definido para cada coleção, com etapas de construção do SIP (metadados, arquivos e termo de submissão), avaliação e conferência de metadados	Política de submissão são construídas através da criação e a atribuição de papéis a grupos de usuários, para dataverses e datasets. Colaborador cria ou atualiza dataset, que é submetido à aprovação para publicação
SUB3	Transferência legal da custódia dos dados ao repositório, autenticação do produtor e usuários envolvidos, aprovação e armazenamento de acordos de submissão, certificação e registros de proveniência.	Fluxo de submissão com autorização e autenticação dos usuários, concordância com o termo de submissão, armazenamento do termo e registro de ações em metadados de proveniência	Submissão realizada por pessoas com papéis autorizados, com a definição de termos (licença). Registro dos envolvidos Registro das versões, com data e pessoa que publicou.
SUB4	Validação e verificação pacotes de submissão (microserviços), verificação e validação de identificadores globais persistentes, geração dos pacotes de armazenamento para submissões aceitas	Fluxo para construção e aprovação do pacote de submissão, direcionado o usuário a produzir um pacote com estrutura adequada.	Dataverse disponibiliza uma interface para construção do pacote de submissão, direcionado o usuário a produzir um pacote com relação a sua estrutura adequada. Pacote é submetido a aprovação. Micro serviços para gerenciar formatos e metadados estruturais de arquivos de dados tabulares e geoespaciais
SUB5	Extração, verificação e/ou produção de metadados técnicos, administrativos e descritivos. Registro em metadados das ações de submissão.	Não extrai metadados técnicos dos formatos. Não verifica formatos. Identifica e descreve o formato de arquivo submetido pela terminação do nome	Não extrai metadados técnicos dos formatos Identifica e verifica formato usando ferramenta JHOVE
SUB6	Submissão em lote. Submissão por máquina e em ambientes distribuídos, recepção de pacotes submetidos por outros ambientes e/ou depósito de materiais aprovados em outros ambientes, via protocolos e estruturas de pacotes de submissão aceitos e padronizados, como Sword, OAI-PMH (alimentação da BD via colheita, BD é agregador), OAI-ORE	Submissão via SWORD Submissão por Colheita via OAI-PMH API Própria Colheita via OAI-ORE	ubmissão via Sword Submissão por Colheita via OAI-PMH API própria Submissão via OJS (Integração com OJS)
PD1	Serviços/microserviços para garantir acesso a longo prazo: Gerenciamento de armazenamento, checagem de integridade e ações de preservação digital, como refrescamento (troca de mídia), replicação (cópias de segurança), reempacotamento (reestruturações de pacotes), transformação (migração).	Módulo de Curadoria Digital, que verifica a ocorrência de vírus e checagem de metadados obrigatórios Exportação dos pacotes, para que a preservação digital seja feita em ambiente externo	Verificações somente na submissão

PD2	Planejamento e ações de preservação digital: gerenciamento de formatos de arquivos (políticas para formatos aceitos), planejamento e execução automatizada de ações planejadas (especificação de planos e execução de microserviços quando objetos submetidos ou armazenados não estão em conformidade com os planos atuais), informações de apoio à preservação digital (estatísticas de uso dos formatos, mídias etc.)	Registra formatos e suas políticas (aceito, aceito mas sem ações de preservação, não aceito)	Não
PD3	Integração com serviços de preservação digital de terceiros ou colaborativos, como Lockss, DuraCloud, armazenamento nas nuvens, etc. Integração com serviços ou infraestruturas locais de preservação digital, como iRods e Archivemática	LOCKS Duracloud Pacotes exportados podem ser importados e gerenciados por Archivematica	Archivematica usa API de Dataverse para buscar dataset, constrói pacote de submissão, e ingere pacote de submissão
PD4	Exclusão de dados, com manutenção dos metadados, autorizações e registro da ação	Excluídos permanecem	Excluidos permanecem
PD5	Usado em Repositório Digital Confiável certificado	Sim	Sim
AC1	Recuperação de informação (busca por metadados, busca por ocorrência de palavras, navegação por facetas (assunto, título, produtores, tipos de dados, unidades, subunidades, grupos, estudos,...)	Busca por metadados. Navegação por facetas	Busca por metadados e facetas
AC2	Informações sobre direitos de uso, licenças	Sim	Sim
AC3	Informações de proveniência e para uso dos dados (metadados descritivos, metadados de proveniência, documentação)	Metadados registram ações do fluxo de submissão	Metadados com informações de proveniência. Metadados de proveniência para arquivos, com a possibilidade de importação em PROV. Registro das versões
AC4	Informações de citação	Exibe handle. Geração de citação pode ser construída na interface de usuário	Gera citação a partir de metadados, incluindo DOI/ Handle e UNF
AC5	Identificadores globais e persistentes para conjuntos de dados, considerando versionamento e objetos digitais em várias representações, e usando serviços, protocolos e padrões aceitos pela comunidade (como os padrões DOI e Handle System e os serviços DataCite, ...)	Identificadores persistentes: Handle ou DOI Acesso via protocolos HTTP e HTTPS	Identificadores persistentes: Handle ou DOI Acesso via protocolos HTTP e HTTPS

AC6	Identificadores globais e persistentes para recursos relacionados aos conjuntos de dados e usando serviços, protocolos e padrões aceitos pela comunidade (como os ORCID para pesquisadores)	Uso do serviço ORCID para controle dos nomes de autoridade	ORCID ou outros esquemas podem ser registrados junto a pessoas cadastradas
AC7	Acesso aos dados e aos metadados via identificador e protocolo aberto	Sim	sim
AC8	Restrições de acesso (nível de arquivo, conjunto de dados, grupos, unidades etc.; a determinados grupos), embargos , acesso mediante registro do usuário, registro de solicitações de uso e relacionamento com usuários (guestbook)	Acesso livre ou acesso restrito a usuários autorizados e autenticados para comunidades, coleções ou itens Permite a definição de políticas de acesso, para comunidades, coleções e item, através da criação de grupos de usuários e da atribuição de autorizações de acesso a esses grupos. Permite também restrição de acesso por motivos de embargo	Datasets e dataverses, quando publicados, são de acesso público. Arquivos de datasets publicados podem ter acesso restrito Acesso livre a datasets, dataverses e arquivos publicados. Arquivos publicados podem ter seu acesso restrito a usuários autorizados e autenticados Livro de Visitas (acesso condicionado a resposta de questionário)
AC9	Gerenciamento e autenticação de usuários envolvidos com acesso	LDAP, Shibboleth, IP Address e certificado X.509	Autenticação via Shibboleth e OAuth2 (Google, Facebook e ORCID)
AC10	Entrega de dados ao consumidor (pessoa ou sistema) nas representações (formatos e estruturas) adequadas, acesso via API	OAI-ORE	API Própria
AC11	Entrega dos metadados ao consumidor (pessoa ou sistema) nas representações (formatos e estruturas) adequadas, protocolos de colheita de metadados (OAI-PMH, Z39-50), acesso VIA API	OAI-PMH	OAI-PMH Exportação/Download
AC12	Ferramentas para visualização e análise de dados	Não	TwoRaven WorldMap
AC13	Estatísticas e relatórios de uso	Downloadas por item	Downloads API de Métricas
AC14	Acesso às descrições em formatos para Linked Data/Web Semântica	Colheita RDF via OAI-PMH Replicação dos metadados em triplestore	Não
AC15	Recuperação em ambiente Linked Data/Web Semântica (endpoints, SPARQL)	Replicação dos metadados em triplestore	Não

Fonte: Dados da pesquisa

Conclusão

Considerando os três *software*, Dataverse possui recursos para configuração de vários tipos de ambientes de repositório, incluindo hierarquias organizacionais e políticas de gestão distintas para unidades ou grupos, incluindo esquemas de metadados e licenças. Isso é possível em DSpace, entretanto, exige adaptações ou configurações, com algumas limitações no controle de versões. CKAN já é mais limitado, entretanto, é uma boa alternativa quando usado como serviço de publicação e de acesso, com a submissão e preservação digital sendo realizadas por outros ambientes.


UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL
FURG
UNIVERSIDADE FEDERAL
DO RIO GRANDE
ibict
RNP
REDE NACIONAL DE
ENSINO E PESQUISA