ACESSO ABERTO A DADOS DE PESQUISA NO BRASIL: PRÁTICAS E SOLUÇÕES TECNOLÓGICAS

Caterina Groposo Pavão[1]; Eduardo Nunes Borges[2]; Leandro Ciuffo[3]; Luís Alberto Barbosa Azambuja[2]; Rafael Port da Rocha[1]; Rene Faustino Gabriel Junior[1]; Samile A. de Souza Vanz[1]; Sônia Elisa Caregnato[1]
[1]UFRGS; [2]FURG; [3]RNP
dadosabertos@rnp.br; cedap@ufrgs.br

Resumo
Governos e instituições passam a identificar o valor estratégico do compartilhamento dos dados de pesquisa e fomentar o desenvolvimento de infraestruturas e tecnologias que estimulam o acesso aberto a dados de pesquisa (AADP). O objetivo deste trabalho é identificar as práticas e soluções tecnológicas de acesso aberto a dados de pesquisa no Brasil, focando especificamente na identificação de práticas de AADP em instituições brasileiras; na identificação de potenciais usuários nacionais de serviços de AADP; e na busca de soluções tecnológicas para repositórios. Para identificar os repositórios brasileiros em AADP, recorreu-se ao diretório RE3data e ao questionário aplicado a 4.735 líderes de grupos de pesquisa do CNPq e coordenadores de programas de pós-graduação. Foram localizados 15 repositórios brasileiros, dos quais quatro com abrangência internacional, envolvendo várias instituições; onze de abrangência nacional, sendo cinco multi-institucionais; predominam as áreas de geociências e ciências biológicas e agrárias; com características de repositórios disciplinares. O questionário "Práticas e percepções sobre acesso aberto a dados de pesquisa no Brasil" apresentou um panorama do AADP no Brasil, dentre os resultados preliminares, apresenta que 68,0% dos pesquisadores compartilharam dados de alguma forma; 39,5% utilizaram dados abertos compartilhados por outros grupos; 31,7% utilizaram algum repositório para acessar dados; 31,9% indicaram a falta de infraestrutura como uma das razões que dificultam o compartilhamento e 21,4% indicam a falta de padronização dos dados para serem compartilhados; 47,7% acham que a instituição do pesquisador deve oferecer serviços de apoio à gestão dos dados de pesquisa. No Brasil, os repositórios utilizam ferramentas tecnológicas variadas, muitas de desenvolvimento próprio para o acesso aos dados de pesquisa. O mais frequentes foram o DSpace, Metacat e Dataverse.
Palavras-chave: Acesso aberto, dados de pesquisa, repositórios de dados de pesquisa.

Introdução

Ao identificar o valor estratégico do compartilhamento dos dados de pesquisa, governos e instituições passaram a fomentar o desenvolvimento de infraestrutura e tecnologias que estimulam o Acesso Aberto a Dados de Pesquisa (AADP). No Brasil, a busca de subsídios para ações de prover AADP envolve a identificação das práticas nas instituições brasileiras, identificação dos potenciais usuários, levantamento e experimentação de serviços e soluções tecnológicas para compartilhamento de dados. Neste contexto, este artigo apresenta resultados parciais obtidos pelo Grupo RDP-Brasil na identificação de práticas e soluções tecnológicas de AADP no cenário brasileiro.

O objetivo deste trabalho é verificar as práticas e soluções tecnológicas de acesso aberto a dados de pesquisa no Brasil. Focando especificamente na identificação de práticas de AADP em instituições brasileiras, os potenciais usuários nacionais de serviços de AADP e as soluções tecnológicas disponíveis para repositórios de dados científicos.

O restante do texto está organizado da seguinte forma. A Seção 1 introduz fundamentos sobre repositórios de dados. A metodologia do estudo é apresentada na Seção 2. Diferentes soluções tecnológicas para repositórios de dados de pesquisa são analisadas na

Seção 3. Na Seção 4 são apresentados os resultados quanto a identificação de usuários de AADP e as práticas da comunidade científica brasileira. Por fim, a Seção 5 conclui e apresenta direções futuras.

1 Repositórios de dados

Os repositórios de dados podem ser categorizados como institucionais, a exemplo do heiDATA da Heidelberg University, data.bris da University of Bristol ou Edinburgh DataShare da universidade alemã de mesmo nome. Também podem ser multidisciplinares, como o Figshare, dataHub, Zenodo entre outros, onde são depositados dados de diversos tipos de pesquisa e áreas do conhecimento. Os repositórios também podem ser disciplinares, onde são depositados e agregados dados de uma área do conhecimento ou especialidade. Como exemplos desses repositórios estão o UK *Data Archive*, que reúne dados do Reino Unido nas áreas de Ciências Sociais e Humanidades; o *Protein Data Bank*, banco de dados para dados estruturais tridimensionais de grandes moléculas biológicas, como proteínas e ácidos nucléicos; e o ICPSR que reúne arquivos de pesquisa nas ciências sociais e comportamentais, com coleções especializadas de dados em educação, envelhecimento, justiça criminal, abuso de substâncias, terrorismo e outros campos.

Em áreas onde os dados são altamente heterogêneos e dispersos entre instituições, projetos, laboratórios, pequenos grupos de pesquisa e pesquisadores individuais, estes dados compõem a cauda longa dos dados de pesquisa. Ou seja, esses dados, usuários, questões de pesquisa e metodologias diferem fundamentalmente da forma como a grande ciência trabalha (SALES; SAYÃO, 2018). Heidorn (2008) destaca que o amplo espectro da distribuição da cauda longa nos sinaliza para o fato que os dados gerados ou coletados em decorrência dos pequenos projetos de pesquisa são distribuídos por todos os domínios do conhecimento, das artes e humanidades até as áreas mais identificadas como os padrões da grande ciência como física e astronomia. Estas pequenas coleções de dados estão sendo reconhecidas como ativos informacionais de alto valor, que coletivamente tem o potencial de ser mais relevante que a soma de suas partes (WYBORN; LEHNERT, 2016).

Os repositórios disciplinares armazenam dados que seguem padrões e recomendações de disciplinas específicas. Muitas dessas áreas possuem padrões e soluções tecnológicas específicas, coordenadas por associações ou organização, como exemplo da Elixir[1] que define padrões de dados para a área de Ciência Biológica na Europa. Já os repositórios institucionais se caracterizam por agregar uma diversidade de tipos de dados de pesquisa, atendendo as mais diversas características, podendo assim serem caracterizados como repositórios de cauda longa dos dados.

Os 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 exemplo da *Data Archiving and Network Services* (DANS) agregando diversas instituições em um serviço denominado DataverseNL, onde as instituições participantes tem espaços para armazenar seus dados de pesquisa; outro exemplo e a Texas Data Repository (TDL), que reúne diversas universidades do Texas nos Estados Unidos.

O desenvolvimento ou criação de repositório deve seguir uma série de modelos de referência e princípios, de forma a manter o máximo de compatibilidade e padronização arquivística, recomenda-se o uso do Modelo de Referência do OAIS - Open Archival

[1] https://www.elixir-europe.org/

Information System, definido pela norma ISO 1472:2003. Além de atender esse modelo, os repositórios devem disponibilizar seus dados seguindo os princípios FAIR e de citação.

O modelo OAIS foi desenvolvido para definir padrões de preservação de informações digitais a longo prazo, de forma a facilitar a homogeneização, independente de disciplina, sobre os requisitos de um arquivo ou repositório. O modelo OAIS fornece estrutura conceitual genérica para a construção de repositório de arquivamento completo e identifica as responsabilidades e interações de Produtores, Consumidores e Gerentes de documentos em papel e digitais (OAIS, 2018).

FAIR é um conjunto mínimo de princípios orientadores e práticas aceitadas pela comunidade para que os produtores e os usuários, humanos ou computadores possam usar mais facilmente os dados e cita-los corretamente. Os princípios FAIR são ligados aos diretrizes de citações de dados definidos na *Joint Declaration Data Citation Principles* (JDDCP) do Force 11, sendo que a sigla FAIR representa (F) achável, (A) Acessível, (I) Interoperável e (R) Reutilizável (WILKINSON; et al., 2016). Os Princípios FAIR enfatizam especificamente o aprimoramento da capacidade das máquinas de encontrar e usar os dados automaticamente, além de apoiar sua reutilização por indivíduos.

De forma a estabelecer um acesso mais fácil e inequívoco aos dados de pesquisa, esses devem utilizar identificadores persistentes (IP), como o *Digital Object Identifier* (DOI) ou Handle. O IP é atribuído por organizações que operam em âmbito mundial, a DataCite reúne instituições de todo o mundo, e aos seus associados possibilita o registro de DOI para dados de pesquisa. A DataCite ajudar a comunidade de pesquisa a localizar, identificar e citar dados de pesquisa com confiança. O DataCite é uma organização global sem fins lucrativos e oferece vários serviços para seus membros de forma a atender às diversas necessidades da comunidade de pesquisa global. Também reúne a comunidade para enfrentar os desafios de tornar os dados da pesquisa visíveis e acessíveis (DATACITE, 2018).

Os repositórios além da infraestrutura tecnológica devem garantir sua perenidade, dessa forma o desenvolvimento de um repositório deve considerar critérios para construção de repositórios digitais confiáveis, que são certificados como tal, garantindo a qualidade e permanência dos depósitos por um longo tempo. Exemplos de agências certificadoras são Data Seal Approval, Core Trust Seal e Nestor, com destaque para a norma ISO 16363:2012 para repositórios confiáveis. As certificações não visam somente o atendimento dos critérios técnicos, mas abrange questões mais amplas como a infraestrutura organizacional, o gerenciamento dos objetos digitais e a infraestrutura tecnológica, técnica e de segurança, de forma a garantir o acesso a um longo prazo com a garantia da manutenção dos serviços.

Além da infraestrutura e políticas os repositórios precisaram de aplicações que possibilitem a concretização. O estudo realizado por Moreno (2018) investigou as características dos dados de pesquisa no cenário espanhol. A autora selecionou repositórios exclusivamente espanhóis registrados no diretório re3data.org a fim de identificar o uso de sistemas de informação/infraestrutura, a tipologia de dados e metadados relacionados, bem como as áreas mais representativas na disponibilização dos dados. De seus resultados, identificou que o DSpace e o Dataverse foram os mais utilizados.

Em outro levantamento realizado por Johnston (2016), a autora buscou nos Estados Unidos softwares para repositórios de dados de pesquisa, como resultado identificou outros como o DigitalCommons, EPrints, Fedora, Nesstar, HubZero, Hydra, Islandora, Zenodo entre outros. Além dos aplicativos citados, O CKAN é apontado por alguns autores como uma ferramenta importante na organização e preservação de dados de pesquisa. Em estudo realizado por Shankar e Bhimrao (2018) que realizaram uma comparação dos softwares CKAN e Dataverse, concluíram que a escolha do aplicativo deve estar relacionado diretamente ao tipo de repositório e que tipo de serviço suporta melhor o repositório de dados.

O estudo parte dos resultados parciais desenvolvidos no mapeamento de potenciais usuários nacionais de serviços AADP no Brasil. Para identificar usuários nacionais de serviços AADP no Brasil foi aplicado um questionário eletrônico para os coordenadores de programas de pós-graduação e de grupos de pesquisa no Brasil no mês de abril de 2018.

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.

Os requisitos para análise das soluções tecnológicas foram estruturados com base no modelo OAIS, adicionados de requisitos relativos ao desenvolvimento e uso software. Os requisitos foram organizados em categorias. Essas categorias foram estruturadas com base no modelo funcional de OAIS e na metodologia de OAIS para identificar a interface entre produtor e a base de dados.

As categorias são: 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.

A pesquisa de identificação das práticas de AADP em instituições brasileiras ocorreu em três etapas, entre março e junho de 2018. Na primeira delas, os dados foram coletados a partir do diretório internacional Research Data Repositories Information (Re3data). Para busca, foi utilizada a opção Browser by country do menu principal. Em seguida, foi selecionada a opção Brasil. Complementando a lista, foram identificados nas respostas do questionário repositórios de dados de pesquisa que não estão catalogados no RE3data. De forma a buscar mais informações sobre o repositório, foram consultados os sites dos repositórios localizados.

Para identificar as ferramentas tecnológicas mais utilizadas, foram utilizados os dados do diretório RE3data em âmbito mundial. O diretório disponibiliza estatísticas sobre vários indicadores, como software, linguagem, temática, entre outros. Os repositórios brasileiros identificados na pesquisa serão caracterizados para identificação da tecnologia utilizada.

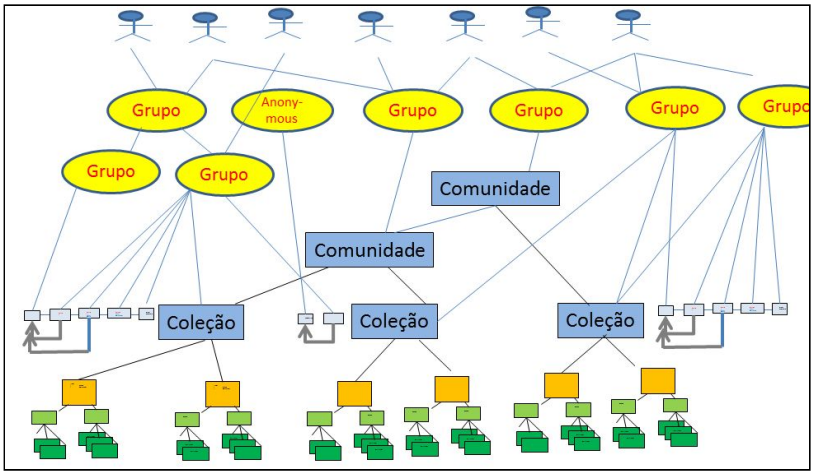
2 Soluções tecnológicas para repositórios de dados de pesquisa

A análise do estudo limitou-se ao três aplicativos, o DSpace, o Dataverse e o CKAN, por esses serem o que apresentam a maior frequência de instâncias catalogados no RE3Data.

O DSpace foi concebido do esforço colaborativo entre a MIT Libraries e a Hewlett-Packard Company com o objetivo na criação de repositórios digitais com funções de armazenamento, gerenciamento, preservação e visibilidade da produção intelectual, possibilitando a adoção de instituições de diversos segmentos, sendo um software livre. O DSpace foi desenvolvido utilizando os princípios do OAIS com uma arquitetura que permite o gerenciamento da produção científica em comunidades e coleções. O software é versátil, integrando qualquer tipo de material digital, como arquivos de áudio e vídeo, páginas web, coleções de bibliotecas digitais, livros, teses, programas de computador, entre outros (MARTINS; SILVA; SIQUEIRA, 2018).

O DSpace possui recursos que permitem variadas configurações para ambientes de repositório de forma a representar entidades organizacionais, como organizações, unidades, subunidades e grupos (Figura 1). As comunidades podem conter subcomunidades e coleções, podendo haver coleções que representam conjuntos de itens.

Figura 1 - Comunidades e Coleções



Fonte: Autores (2018).

O DSpace não foi desenvolvido para dados de pesquisa, entretanto, comunidades podem ser usadas para representar instituições produtoras ou custiodadoras de conjuntos de dados, e/ou grupos de pesquisa. Por exemplo a Universidade de Edinburg utiliza em seu repositório de dados de pesquisa, Edinburgh DataShare, comunidades e coleções para representar hierarquicamente a comunidade de pesquisa da Universidade.

DSpace dispõe de recursos para implementar comunidades e coleções com políticas de funcionamento próprias e distintas, possibilitando a criação de fluxos de submissão a publicação. A ferramenta ainda possibilita o registro e autenticação de usuários; a organização de usuários em grupos; e a atribuição de autorizações para grupos e usuários com a definição de políticas em que somente grupos autorizados podem acessar comunidades, coleções e itens. 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.

DSpace não foi desenvolvido para ser um *Data Research and Information Management* (DRIM). Permite somente a representação de resultados de pesquisa documentais (textos, planilhas, imagens, etc). No DSpace não é possível a representação de recursos de DRIMs, como projetos, pessoas/pesquisadores, unidades organizacionais, equipamentos, laboratórios, etc. Unidades organizacionais podem ser representadas em DSpace por meio de comunidades, entretanto, estas são somente vistas pelo *software* como gestoras e contendedoras de conjuntos de dados, não dispondo de propriedades para descrição de unidades organizacionais, nem propriedades que as relacionam com outros recursos de um Sistema Integrado de Pesquisa (SIP).

De forma a ampliar as relações de pesquisa foi desenvolvido o DSpace-CRIS, sendo a primeira extensão gratuita de código aberto do DSpace para DRIM já desenvolvido. O DSpace-CRIS permite a representação e a interligação de pesquisadores, unidades organizacionais, projetos (*grants*) e resultados de pesquisa (publicações, patentes, teses). Possibilita ainda 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 do

Common European Research Information Format (CERIF) para representação de DRIMs. Algumas instâncias dos softwares DSpace-Cris podem ser visualizadas na University of Hong Kong (HKU). Nesse repositório, os conjuntos de dados são representados por meio de classe Dataset. Essa classe contém propriedades que descrevem esses conjuntos e que estabelecem relações com outros objetos do SIP e arquivos.

Desde a versão 5 o DSpace incorporou um módulo de integração com a Web Semântica, possibilitando por meio desse módulo, converter o conteúdo armazenado em Dspace em triplas *Resource Description Framework* (RDF), 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.

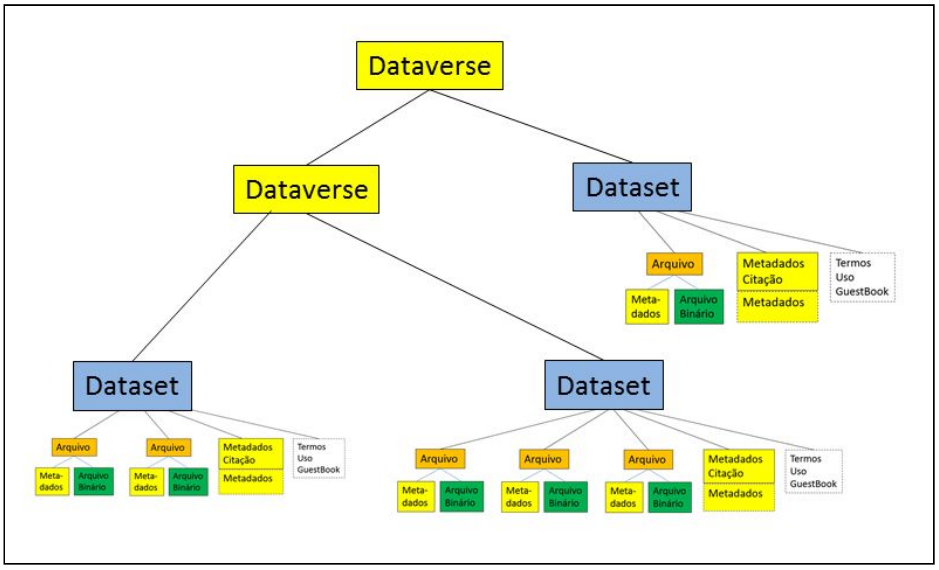
O Dspace são não possui recursos para versionamento de dados de pesquisa. Entretanto, uma extensão de DSpace foi desenvolvida pelo repositório Dryad para prover serviço de versionamento. No Dryad, um estudo é representado como um item do DSpace, o que possibilita que cada conjunto de dados seja atribuído seu próprio identificador global persistente (DOI e Handle). Em Dryad, a atualização de algum arquivo leva a uma nova versão, e cada item dessa versão (cada arquivo) possui um novo identificador.

O DSpace permite a definição de esquemas de metadados com elementos não hierárquicos e incorporação de Perfis de Aplicação para cada tipo de coleção. Dessa forma, DSpace possibilita a definição de esquemas de metadados que atendam às necessidades da comunidade do repositório. Entretanto, DSpace não permite a criação de elementos com estruturas complexas, isto é, elementos cujos conteúdos são também desdobrados em novos elementos. Isso traz dificuldades na configuração da ferramenta para usar esquemas de metadados que são especificados em XML e que contém elementos em estrutura hierárquica, como DDI.

Da mesma forma que o DSpace, o Dataverse foi desenvolvido em Java com o uso do TomCat7 ou Goldfish no ambiente Web. O Dataverse é uma plataforma tecnológica de código aberto desenvolvida pela Equipe Dataverse no Instituto de Ciências Sociais Quantitativas (IQSS) da Universidade de Harvard (DATAVERSE, 2018).

O Dataverse é um aplicativo da Web de código aberto para compartilhar, preservar, citar, explorar e analisar dados de pesquisa (CROSAS, 2011). Suas funcionalidades de coleta e de exportação de metadados permitem a circulação de grandes quantidades de dados, expandindo seu arquivamento e preservação (KING, 2007).

Figura 2 - Representação do Ambiente do Repositório



Fonte: autores (2018).

A Figura 2 apresenta as possibilidades de organização das comunidades (Dataverse) e das coleções (Dataset) dentro do software. Destaca-se que podem ser hierarquizados Dataverses dentro de outros Dataverses, facilitando seu gerenciamento.

O Dataverse, assim como o DSpace, foi desenvolvido de forma a atender os critérios estabelecidos pelo OAIS, sendo uma solução tecnológica voltada para a gestão de dados pesquisas científicas, onde deve disponibilizar entre suas funções a atribuição de identificadores persistentes, interface multilíngue para sua comunidade de usuários, fácil recuperação, e possibilidades de interoperabilidade atendendo aos princípios FAIR.

Dentro dos princípios FAIR o Dataverse é a sua capacidade de incluir metadados de citação para cada conjunto de dados submetido. Os metadados de citação no software atendem a aos critérios de citação, e são compatíveis com DataCite, Dublin Core, DDI (*Data Documentation Initiative*) (ALTMAN; CROSAS, 2013).

Uma das principais características do Dataverse está na gestão de depósitos, o software possibilita após a criação dos "Dataverses" processar formatos e descrição de conjuntos de dados e arquivos diferenciados, fazendo com que esses arquivos sejam facilmente recuperáveis e reusáveis, com conversão de arquivos em formatos acessíveis e interoperáveis, inclusive os arquivos de dados tabulares, cuja exploração no Dataverse deve ser especificada.

Tavares, Arellano e Nakagomi (2018) destacam que para um bom funcionamento de DRIM, estes devem informar aos pesquisadores, curadores e instituições representados na plataforma orientações básicas para seu uso, bem como ter a estrutura hierárquica das coleções definida por uma política de gestão de dados de pesquisa que descreva as áreas, projetos e o tipo de materiais que o sistema contemplará.

No Dataverse, dastaca-se também, o uso de *Application programming interface* (API), que possibilita a utilização de aplicações abertas para pesquisar, depositar e acessar dados e visualizar dados armazenados.

O CKAN é uma ferramenta para a construção de portais de dados abertos, que auxilia na publicação e no gerenciamento de coleções de dados, sendo uma plataforma *open source* (CONEGLIAN; SEGUNDO, 2016). O CKAN foi desenvolvido Open Knowledge Foundation, que tem como diretrizes tratar do desenvolvimento, do suporte e da promoção de ferramentas que busquem facilitar a criação, o acesso e a disseminação de conhecimentos (WINN, 2013).

Os dados são o produto principal do CKAN, sendo necessário que exista uma organização de como os dados devem estar estruturados. Da mesma forma que o Dataverse e o DSpace, o CKAN organiza todos os dados em datasets (conjuntos de dados), sendo que cada dataset contém informações acerca de um determinado contexto, como por exemplo informações governamentais de gastos com obras públicas (CONEGLIAN; SEGUNDO, 2016).

No CKAN, a partir da organização dos datasets, os usuários podem fazer buscas e receber informações individualmente, e não todo o conjunto (CONEGLIAN; SEGUNDO, 2016). Basicamente, um dataset contém dois tipos de informações: os metadados que descrevem informações dos dados, como data de criação, criador, título e; conjuntos de recursos, que são os dados em si, podendo ser os recursos de qualquer formato, como PDF, Comma-Separated Values (CSV).

Dentro os diversos benefícios no uso no CKAN, Santarem Segundo e Faria (2013) destacam que o CKAN oferece a possibilidade de *harvesting*, utilização de diversos conceitos de publicação da informação em ambientes abertos, sendo que o CKAN é atualmente uma opção mais adequada para a publicação de dados em formato aberto disponível para uso.

Dentro dos elementos de metadados o utilizado para a representação dos objetos digitais, estão representados: identificador único; descrição; histórico de revisão; visualização de dados; campos extras; licença; tags; grupos e; múltiplos formatos (CKAN, 2015).

3 Identificação de usuários de AADP

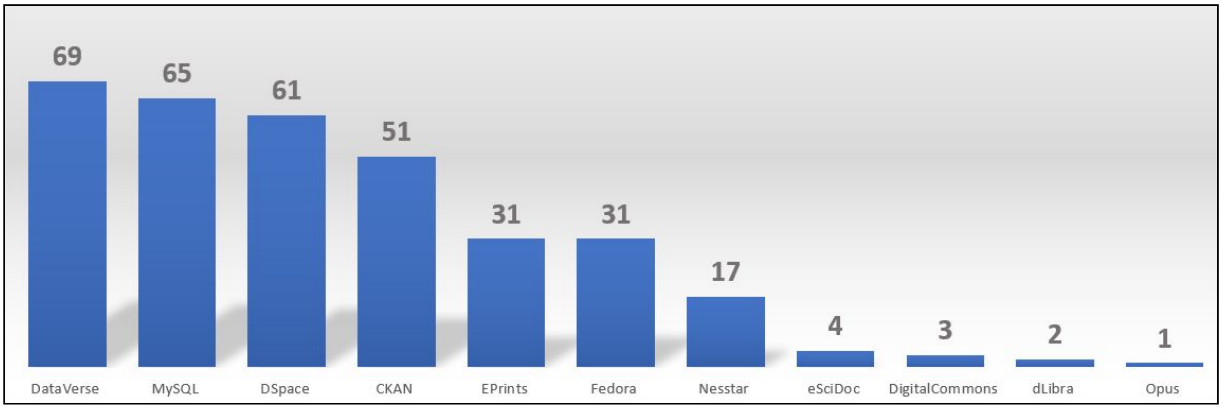
O questionário “Práticas e percepções sobre acesso aberto a dados de pesquisa no Brasil” teve um alcance de 4.735 resposta. Dentre os resultados preliminares foi possível observar que 68,0% dos pesquisadores responderam que compartilharam dados de alguma forma, sejam disponibilizados dados em repositórios, ou compartilhando dados com seus pares diretamente; e 39,5% utilizaram dados abertos compartilhados por outros grupos. Números esses considerados relevantes para o desenvolvimento de projetos de repositórios de dados de pesquisa, visto que apenas 31,7% já utilizaram algum repositório para acessar dados.

No questionário pode-se perceber que a falta de infraestrutura foi apontada por 31,9% das respostas, como razões que dificultam o compartilhamento e 21,4% indicam a falta de padronização dos dados. Quase metade dos respondentes, 47,7% acham que a instituição do pesquisador deve oferecer serviços de apoio à gestão dos dados de pesquisa.

Uma das conclusões do questionário demonstra que pesquisadores que já consumiram dados de pesquisa, tendem a compartilhar os seus dados e pesquisadores que nunca consumiram dados, mas que conhecem gestão de dados, também compartilham seus dados.

No estudo também pode observar que a maioria dos pesquisadores nunca compartilhou dados de pesquisa e que 25,0% já elaboraram um plano de gestão de dados. Demonstrando ainda que existe um grande percurso na ciência brasileira para os dados de pesquisa.

Figura 3 – Softwares utilizados nos Repositórios registrados no RE2data



Fonte: Re2data (2018)

No Re3data foram identificados um total de 2.613 repositórios, dos quais 1.842 (70,5%) são disciplinares, 557 (21,3%) institucionais e 214 (8,2%) não definidos. Destaca-se ainda que a 1.022 dos repositórios encontra-se nos Estados Unidos, 341 na Alemanha e 269 no Reino Unido. Sobre o acesso aberto, pode-se identificar que 55,7% mantem os dados em acesso aberto, e 30,7% em acesso restrito, sendo o restante 13,7% ficam fechados ou estão embargados.

Na busca de soluções tecnológicas, o DSpace, CKAN e Dataverse são os softwares mais usados para repositórios de dados (Figura 3). DSpace e Dataverse são soluções completas e integradas, baseadas no modelo OAIS e usadas por repositórios certificados como confiáveis (DSA/Core Trusted Seal). Dispõem de funções para organização e configuração de coleções em comunidades, processos de submissão por comunidades, uso de diversos esquemas de metadados, identificador de objeto persistente, etc. Dataverse foi desenvolvido especificamente para dados de pesquisa, trazendo facilidades como metadados de citação e versionamento. CKAN foi desenvolvido para dados abertos, mas possui limitações para a implantação de processos de submissão e representação de esquemas de metadados. É solução indicada para serviço de acesso a dados em que submissão e a preservação são realizadas por outros softwares. Além de soluções completas e integradas, repositórios também são desenvolvidos a partir de serviços independentes, pelo reuso de diversos tipos de software, como EUDAT. Essa solução oferece maior flexibilidade de adaptação às necessidades de repositório com características especiais para coleta automática e representação de dados.

Um dos softwares identificados no RE3data refere-se ao MySQL, que não se trata de um Repositório de Dados de Pesquisa, mas um Sistema de Gerenciamento de Banco de Dados (SGBD) que armazena dados, pode-se justificar essa frequência na categorização pela facilidade de armazenar dados, em sistemas desenvolvidos pelos pesquisadores ou pela instituição.

Ao filtrar os repositórios Brasileiros no RE3data, identifica-se um total de sete repositórios, comparando esses resultados com o do questionário, pode-se identificar outros 8 repositórios na catalogados no RE3data. O quadro 1 apresenta os repositórios e suas tecnologias.

Quadro 1 – Repositórios de dados de pesquisa brasileiros

#	**Repositório**	**Instituição**	**Software**
1	IBICT Dataverse Network	Ibict	Dataverse
2	Banco de Dados de Exploração e Produção (BDEP)	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) e Serviço Geológico do Brasil.	Desconhecido

3	Global Collaboration Engine (GLOBE)	Instituto Nacional de Pesquisas Espaciais (INPE)	Desconhecido
4	International Ocean Discovery Program (IODP)	Integrated Ocean Discovery Program (IODP)	Desconhecido
5	PPBio Data Repository	s Centro de Estudos Integrados da Biodiversidade Amazônica (INCT-CENBAM)	Metacat
6	Global Climate Data (WorldClim)	s Centro de Referência em Informação Ambiental (CRIA)	Drupal
7	Base de Dados Científicos da Universidade Federal do Paraná (BDC/UFPR)	Universidade Federal do Paraná	DSpace
8	Base Tuiuiu	Empresa Brasileira de Pesquisa Agropecuária (Embrapa),	Desenvolvimento local
9	Consórcio de Informações Sociais (CIS),	Universidade de São Paulo (USP)	Desenvolvimento local
10	Instituto Nacional de Meteorologia (INMET)	Ministério da Agricultura Pecuária e Abastecimento (MAPA), Brasil	Desconhecido
11	Instituto Brasileiro de Geografia e Estatística (IBGE)	Instituto Brasileiro de Geografia e Estatística (IBGE), Brasil.	Desconhecido
12	Sistema Maxwell	Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Brasil.	Desenvolvimento local
13	Repositório de dados PELD	Ministério de Ciência, Tecnologia, Inovações e Comunicações (MCTIC)	Metacat
14	Projeto speciesLink	Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)	SpLinker
15	IPAgriDados	Empresa Brasileira de Pesquisa Agropecuária (Embrapa), Brasil.	DSpace

Fonte: Pavão (2018).

A partir dos 15 repositórios brasileiros identificados observaram-se as seguintes práticas: quatro deles são de abrangência internacional, envolvendo várias instituições; dos sete de abrangência nacional, cinco são multi-institucionais; predominam as áreas de geociências e ciências biológicas e agrárias; somente cinco divulgam informações sobre políticas e quadro apresentam os padrões de metadados usados (PAVÃO, 2018).

Observa-se também que no Brasil existem iniciativas de desenvolvimento de repositórios de dados de pesquisa, entretanto utilizando especificidades da área disciplinar ou da própria instituição, não aplicando os princípios do OAIS e do FAIR. Dois repositórios utilizam o DSpace com ferramenta tecnológica, e apenas um utiliza o Dataverse. O Metacat não foi localizado na literatura pesquisada, mas é uma ferramenta da DataOne (2018), de livre acesso, para o desenvolvimento de catálogo de metadados e repositório de dados flexível e de código aberto para dados científicos, especialmente da ecologia e da ciência ambiental.

O Metacat utiliza XML como uma sintaxe comum para representar o grande número de padrões de conteúdo de metadados que são relevantes para ecologia e outras ciências. Assim, o Metacat é um banco de dados XML genérico que permite armazenamento, consulta e recuperação de documentos XML arbitrários sem conhecimento prévio do esquema XML. O Metacat está sendo usado extensivamente em todo o mundo para gerenciar dados ambientais.

Considerações Finais

A pesquisa foi baseada na identificação de repositórios de dados de pesquisa no diretório RE3data e nas respostas do questionário aplicado por Pavão e outros pesquisadores

(2018) sobre as práticas de uso e compartilhamento de dados de pesquisa. Pode-se observar que o tema é relevante dentro da comunidade científica. O questionário possibilitou analisar se os pesquisadores depositam seus dados de pesquisa, e onde os fazem.

Observou-se que os pesquisadores têm uma preocupação e interesse no compartilhamento de dados de pesquisa, porém destacam a necessidade de apoio institucional e capacitação de forma a não onerar o pesquisador.

No Brasil pode-se observar que o número de repositórios de dados de pesquisa ainda é baixo, quando comparado as Estados Unidos e Europa. Foram identificados 15 repositórios no diretório RE3data e nas respostas dos questionários. A maioria dos repositórios brasileiros são institucionais, tendo quatro com colaboração com outros países. Das ferramentas tecnológicas, o DSpace e o Metacat foram os mais representativos com duas instalações, entretanto, existe a prevalência no desenvolvimento de aplicações próprias dentro das instituições de forma a atender especificidades da área.

Agradecimentos

O presente trabalho foi realizado com apoio da Rede Nacional de Pesquisa (RNP) – Código de Financiamento Nº 002980/2018.

Referências

ALTMAN, M.: CROSAS, M. The evolution of Data Citation: from principles to implementation. **IASSIST Quarterly**, 2013. Disponível em: <http://www.iassistdata.org/sites/default/files/iqvol371_4_altman.pdf>. Acesso em: 15 out.

CKAN. About. 2015a. Disponível em: <http://ckan.org/about/>. Acesso em: 03 jun. 2016.
CONEGLIAN, C. S.; SEGUNDO, J. E. S. Profissional da informação no contexto dos dados abertos: o uso do ckan para a disponibilização e a organização de dados. **Informação@Profissões**, v. 5, n. 2, p. 55-78, 2016. DOI: 10.5433/2317-4390.2016v5n2p55

CROSAS, M. The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data. **D-Lib Magazine**, v. 17, n.1/2, 2011. DOI:10.1045/january2011-crosas.

DATACITE. Welcome datacite. Disponível em: <https://www.datacite.org/>. Acesso em 10 out. 2018.

DATAONE. Metacat. Disponível em: <https://www.dataone.org/software-tools/metacat>. Acesso em 10 out. 2018.

DSPACE. About Dspace. Disponível em: < https://duraspace.org/dspace/about/ >. Acesso em 10 out. 2018.

JOHNSTON, L. Data Repositories: The Answer that Actually Came with a Question. Librarian e-Science Symposium, 2016. Disponível em: <https://escholarship.umassmed.edu/cgi/viewcontent.cgi?referer=https://www.google.com.br/&httpsredir=1&article=1148&context=escience_symposium>. Acesso em 10 out. 2018.

KING, G. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. **Sociological Methods and Research**, v. 36, n. 2, p. 173-199, 2007.

MARTINS, D. L.; SILVA, M. F.; SIQUEIRA, J. Comparação entre sistemas para criação de acervos digitais: análise dos softwares livres dspace, eprints, fedora, greenstone e islandora a partir de novas dimensões analíticas. **InCID**: Revista de Ciência da Informação e Documentação, v. 9 n. 1, n. 1, p. 52-71, 2018. DOI: 10.11606/issn.2178-2075.v9i1p52-71

MORENO, F. P. Repositórios de dados de pesquisa na espanha: breve análise. Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, v. 23, n. 53, p. 52-63, 2018. DOI: 10.5007/1518-2924.2018v23n53p52

OAIS. OAIS Reference Model (ISO 14721). Disponível em: <http://www.oais.info/>. Acesso em 10 out. 2018.

PAVÃO, Caterina Groposo; VANZ, Samile Andrea de Souza; PASSOS, Paula Caroline Schifino Jardim; CAREGNATO, Sônia Elisa; AZAMBUJA, Luís Alberto Barbosa; BORGES, Nunes Borges; GABRIEL JUNIOR. Rene Faustino; ROCHA. Rafael Port da. Acesso aberto a dados de pesquisa no Brasil: repositórios brasileiros de dados de pesquisa: relatório 2018. Handle: 2050011959/20180801. Disponível em: <http://hdl.handle.net/20.500.11959/127>

SHANKAR, M. S.; BHIMRAO, G. S. A Comparative study of open source data repository software: Dataverse and CKAN. Library Herald, v. 56, n. 1, 2018. DOI : 10.5958/0976-2469.2018.00005.2

TAVARES, M. F. D.; ARELLANO, M. M.; NAKAGOMI, B. Brasília e a memória em registros digitais: traços da paisagem e a preservação de dados. **Revista Ibero-Americana de Ciência da Informação**, v. 11 No 1, n. 1, p. 183-199, 2018. DOI: 10.26512/rici.v11.n1.2018.8474

WILKINSON, M. D. el al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**, 3:160018, 2016. DOI: 10.1038/sdata.2016.18

WINN, J. et al. Open data and the academy: An evaluation of CKAN for research data management. 2013. Disponível em: <http://eprints.lincoln.ac.uk/9778/1/CKANEvaluation.pdf>. Acesso em: 04 out. 2016.
WYBORN, L.; LEHNERT, K. Exploiting the long tail of scientific data: Making small data BIG. In: ERESEARCH AUSTRAULASIA CONFERENCE, Melbourne, Australia, 10-14 Oct. 2016. **Anais...** Melbourne, Austrália, 2016. Disponível em: <https://eresearchau.files.wordpress.com/2016/03/eresau2016_paper_88.pdf>. Acesso em: 10 out. 2018.