

CARACTERÍSTICAS DOS REPOSITÓRIOS DE DADOS CIENTÍFICOS NO BRASIL

Lucas Nóbrega Paganine

Pós-graduado lato sensu em Gestão Pública pela Faculdade Unyleya. Bolsista pesquisador do Programa de Capacitação Institucional – modalidade Desenvolvimento/categoria D (PCI-D/D) no Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT).
lnpaganine@hotmail.com
<https://orcid.org/0000-0001-8978-4742>

Bianca Amaro

Doutora em Linguística Aplicada pela Universidade Pompeu Fabra (UPF) – Espanha. Coordenadora do Programa Brasileiro de Acesso Aberto no Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Ganhadora do prêmio internacional Electronic Publishing Trust for Development (EPT 2015).
bianca@ibict.br
<https://orcid.org/0000-0002-4703-8992>

RESUMO

O objetivo do estudo em questão é observar e descrever os repositórios de dados científicos no Brasil. Para tanto, foi realizada uma análise descritiva, iniciada a partir de um levantamento na plataforma “re3data”. Esse levantamento resultou nos nove repositórios de dados científicos do Brasil, ali cadastrados. A análise foi orientada pela metodologia desenvolvida, adaptada a partir das características essenciais de um repositório científico. Estas foram adaptadas para a realidade dos dados de pesquisa e, então, organizadas em categorias, entre elas, o Armazenamento, a Descrição e a Apresentação dos itens. O baixo número de repositórios, a não padronização e a inconformidade com padrões internacionais estabelecidos são indicativos de que a realidade Brasileira ainda não está preparada, adequadamente, para lidar com as questões que vêm surgindo.

Palavras-chave: Repositório de dados. Dados de pesquisa. re3data.

CHARACTERISTICS OF SCIENTIFIC DATA REPOSITORIES IN BRAZIL

ABSTRACT

The aim of the study is to observe and describe the repositories of scientific data in Brazil. For this purpose, a descriptive analysis was carried out, starting from a survey on the re3data platform. This survey resulted in the nine scientific data repositories in Brazil registered there. The analysis was guided by the developed methodology, adapted from the essential characteristics of a scientific repository. The characteristics were adapted to the reality of the research data and organized into categories (Storage, Description and Presentation of the items). The low number of repositories, non-standardization and non-compliance with established international standards are indicative that the Brazilian reality is not yet adequately prepared to deal with the issues arising.

Keywords: Data repository. Research data. re3data.

Recebido em: 21/03/2020

Aceito em: 30/07/2020

Publicado em: 11/09/2020

1 INTRODUÇÃO

Com o surgimento das novas tecnologias da informação e comunicação (TIC), novas formas como as pessoas interagem e se comunicam, foram desencadeadas. Por consequência, essas mudanças têm gerado, também, significativos processos de transformação na comunicação científica. Nesse cenário de transformações, o movimento

Open Access surgiu apresentando novas soluções, em contrapartida aos problemas desenvolvidos, devido ao sistema de mercado tradicional dos periódicos científicos comerciais.

Sarmiento *et al.* (2005, p. 8) esclarece que “existem três declarações que regem o movimento Acesso Livre, ou seja, as declarações de Budapeste, Bethesda e Berlim. As três apoiam o acesso global e irrestrito ao conhecimento científico” e são descritas, em linhas gerais, por Costa e Leite (2015, p. 3) ao dizer que elas “reuniram iniciativas de sistemas de informação e procedimentos já existentes, que promoviam o acesso amplo às publicações científicas”. Dessa forma, as declarações apresentaram orientações para o desenvolvimento de novos sistemas e ficaram conhecidos como fundadores do Movimento de Acesso Aberto.

Dentre as iniciativas que formalizam e orientam o movimento *Open Access*, destaca-se, aqui, a declaração de Bethesda, que evidenciou, ainda em 2003, a importância do acesso aberto para as ciências da saúde (SARMENTO; *et al.*, 2005). A declaração de Bethesda formalizou a demanda para o tratamento dos dados de pesquisa, no contexto do acesso aberto.

Os repositórios de dados de pesquisa têm sido sistematicamente propostos pela literatura, como a ferramenta adequada para o compartilhamento desses dados de pesquisa. No entanto ainda se encontram diversos desafios, como a padronização dos sistemas e a curadoria dos dados. Ao se considerar o contexto apresentado e a relevância e atualidade do tema, este estudo busca identificar a situação atual de desenvolvimento dos repositórios de dados de pesquisa, no Brasil.

2 METODOLOGIA

Para execução da análise proposta, foi realizada uma pesquisa de levantamento e análise descritiva, que, segundo Gil (2011, p. 28) “têm como objetivo primordial a descrição das características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis”, acerca dos repositórios de dados no Brasil, registrados no re3data. O levantamento dos dados ocorreu em março, do ano 2018.

A análise metodológica desenvolvida foi sintetizada e adaptada, a partir das três características essenciais para um repositório de dados, proposta por Rodrigues *et al.* (2010). Essas características foram utilizadas como categorias de análise dos dados

coletados. Portanto, os resultados foram sistematizados em três grandes conjuntos: armazenamento, descrição e apresentação dos itens, como ilustrado pelo quadro 2, a seguir.

Quadro 1– Categorias e itens analisados.

Categorias	Elementos analisados
Armazenamento	<i>Software</i> : programa(s) utilizado(s) no repositório.
	Sistema de preservação: métodos ou práticas de preservação utilizadas.
	Licença: tipo de licença(s) utilizada.
Descrição	Metadados (padrão): qual padrão de metadados utilizado.
Conclusão	
Categorias	Elementos analisados
Apresentação dos itens	Tipo de conteúdo: caracterização dos dados apresentados no repositório, como por exemplo, dados audiovisuais, textuais, etc.
	Quantidade de itens: número de entradas de itens no repositório.
	Relação do item com o conjunto: se há indicação da relação do dado com o todo.
	Relação do conjunto com a publicação: se há indicação da relação do conjunto de dados com a publicação, onde se encontra.?

Fonte: Adaptado de Rodrigues et al. (2010).

As adaptações da metodologia foram realizadas com o objetivo de sintetizar as tabelas e as adaptar à realidade dos repositórios brasileiros. Nesse sentido, foram retirados os quesitos Repositório aberto e URL persistente da categoria Armazenamento. Por fim, os elementos “Relação do item com o conjunto” e “Relação do conjunto com a publicação” foram conjugados.

3 RESULTADOS E DISCUSSÃO

Partindo da metodologia estabelecida, foram recuperados 9 repositórios, entre as áreas: Ciências humanas e sociais, Geociências, Ciências da vida e Engenharias. Pode-se perceber que os temas estão distribuídos de forma homogênea, exceto entre as Engenharias, em que aparece em apenas um repositório.

Quadro 2 – Repositórios de dados de pesquisa no Brasil

Nome do repositório	Área	Acesso aberto?	Breve descrição
<i>Brazil Exploration and Production Database</i>	Geociências	Sim	Inaugurado em 2000, armazena, organiza e disponibiliza informações geofísicas, geológicas e geoquímicas, além de processar e analisar.
<i>World Clim – Global Climate Data</i>	Geociências	Sim	Apresenta grades e conjuntos de camadas climáticas para mapeamento e modelagem espacial.
GLOBE	Ciências da vida	Sim	O <i>Global Collaboration Engine</i> é um ambiente <i>online</i> colaborativo voltado para variação ambiental, integrando estudos locais e regionais a dados globais
<i>IBICT Dataverse Network</i>	Ciências Humanas e Sociais	Sim	Trata de preservação de longo prazo e de boas práticas de arquivamento, voltado para os participantes da rede Cariniana.
<i>International Ocean Discovery Program</i>	Geociências	Sim	O <i>International Ocean Discovery Program</i> (IODP) é uma colaboração internacional de pesquisa marinha, que recupera dados de sedimentos e rochas do fundo do mar e monitora ambientes subterrâneos.
Conclusão			
Base de Dados Científicos da Universidade Federal do Paraná	Ciências Humanas e Sociais	Sim	Reúne os dados científicos utilizados nas pesquisas publicadas pela comunidade da UFPR em teses, dissertações, artigos de periódicos e outros materiais bibliográficos
<i>PPBio Data Repository</i>	Ciências da vida	Sim	O Programa de Pesquisa em Biodiversidade (PPBio) foi criado em 2004, integrando atividades de pesquisa e disseminando resultados em diversas finalidades, incluindo gestão ambiental e educação. Seu repositório trabalha com dados ecológicos complexos.

Continuação

Nome do repositório	Área	Acesso aberto?	Breve descrição
Conclusão			
<i>CEDAP Research Data Repository – research data</i>	Multidisciplinar	Sim	O CEDAP da UFRGS objetiva reunir os dados científicos utilizados na pesquisa, nas diversas áreas do conhecimento. O repositório visa reunir os dados, com a documentação, a fim de proporcionar um ambiente de estudo das metodologias de uso e reutilização dos dados. É mantida em parceria com o CPD da UFRGS.
<i>Open Research Data @PUC-Rio</i>	Engenharias	Sim	É um agregador para facilitar o acesso aos dados de nova pesquisa, entre muitos outros conteúdos digitais no Repositório Maxwell.

Fonte: Adaptado de RE3DATA.

Um fator relevante a ser destacado é que os únicos repositórios, exclusivamente, brasileiros são os *Exploration and Production Data Bank*, Base de Dados Científicos da Universidade Federal do Paraná, *Open Research Data @PUC-Rio* e o *CEDAP Research Data Repository – research data*. Configurando, apenas, pouco menos do que metade dos repositórios cadastrados. O *IBICT Dataverse Network* e *PPBio Data Repository* são repositórios que, de acordo com o RE3DATA, também estão inclusos na categoria internacional. O GLOBE é uma iniciativa em conjunto com os Estados Unidos, enquanto o IODP, é uma colaboração de diversos países.

3.1 Armazenamento

A categoria inicial de análise, proposta pela metodologia, descreve o armazenamento dos dados de pesquisa, a partir dos aspectos referentes a: a) software utilizado pelo repositório; b) existência de sistemas de preservação e c) tipos de licenças utilizadas de acesso ao conteúdo. Vale destacar que todos os repositórios estudados são de acesso aberto.

- a) *Software* – As informações acerca dos *softwares* utilizados foram estruturadas no quadro 3, a seguir, permitindo-se perceber que, no momento, não há uma predominância de escolha ao se tratar do assunto. Isso inclui, até mesmo,

o desenvolvimento de programas próprios e específicos, como o caso do World Clim e do GLOBE.

Quadro 3 – Armazenamento: Software

Nome do repositório	Nome do software
Base de Dados Científicos da Universidade Federal do Paraná	<i>Dspace</i>
<i>CEDAP Research Data Repository – research data</i>	
<i>PPBio Data Repository</i>	<i>Metacat</i>
<i>World Clim – Global Climate Data</i>	<i>World Clim Version 2.0</i>
<i>GLOBE</i>	<i>ISEA3H Level 12 DGG, generated using DGGRID software version 3.1</i>
<i>IBICT Dataverse Network</i>	<i>Data Verse</i>
<i>International Ocean Discovery Program</i>	Não informado
<i>Exploration and Production Data Bank</i>	
<i>Open Research Data @PUC-Rio</i>	<i>Maxwell System</i> (próprio)

Fonte: Elaboração própria.

Dentre os *softwares* encontrados, o *Dspace*, o *Metacat* e o *DataVerse* merecem uma atenção especial. O *Dspace* é um *software* livre, voltado para a criação de repositórios, realizando armazenamento, gerenciamento, preservação e visibilidade. O *Metacat* é um repositório de dados flexível e de código aberto, voltado para dados da ecologia e ciências ambientais. E o *Dataverse*, também de código-fonte aberto, foi desenvolvido em *Harvard* para compartilhar, preservar, citar, explorar e analisar dados de pesquisa de uma forma geral. Cada repositório é uma instalação que, por sua vez, abriga diversos *dataverses*.

- b) Sistemas de preservação – Com relação aos sistemas de preservação, os repositórios que informaram adotar algum tipo qualquer de medida, foram:
- *IBICT Dataverse Network* com o *Harvard Dataverse Preservation Policy*, que estabelece termos e datas de *backups* e procedimentos para arquivamento digital.
 - *International Ocean Discovery Program*, em seu *ODP Sample Distribution, Data Distribution, and Publications Policy*, que define a preservação

como uma das obrigações do programa, visando o arquivamento para análises futuras. Ele apresenta, também, a forma como esse arquivamento era realizado, antes de 1997 e como é hoje, reiterando que a preservação é um dos requisitos para uma estratégia de sucesso de amostragem.

- *CEDAP Research Data Repository – research*: determina a data que declara apenas oferecer serviço de repositório que inclui preservação digital; e
- *Open Research Data @PUC-Rio* que declara ser membro da Cooperativa Meta Archive (<https://www.metaarchive.org/>) e seus dados estão sendo preparados para serem incluídos no processo de preservação.

Vale destacar que, dentre os repositórios analisados, apenas a Base de Dados Científicos da Universidade Federal do Paraná e o *CEDAP Research Data Repository* não disponibilizam uma política de dados.

- c) Licença – Percebe-se uma preferência pela licença *Creative Commons* (licenças que facilitam o compartilhamento e reuso de forma menos restritiva), mais adequadas para ambientes de acesso aberto. Logo após, temos a *Copyright* (direito autoral que impede exploração, sem prévia autorização) combinada com outro fator, como por exemplo, licenciamento ou termos de uso.

Quadro 4 – Armazenamento: Licença

Nome do repositório	Licença utilizada
<i>World Clim – Global Climate Data</i>	<i>Copyright & Licensing</i>
Base de Dados Científicos da Universidade Federal do Paraná	CC
<i>CEDAP Research Data Repository</i>	
<i>Open Research Data @PUC-Rio</i>	
<i>PPBio Data Repository</i>	
<i>Exploration and Production Data Bank</i>	<i>Copyright & Terms of use</i>
<i>IBICT Dataverse Network</i>	
<i>GLOBE</i>	<i>Public Domain</i>
<i>International Ocean Discovery Program</i>	<i>CC BY</i>

Fonte: Elaboração própria.

Sobre a categoria armazenamento, percebe-se, então, que não há um *software* predominante utilizado. Assim, a preservação não recebe a preocupação devida e não existe consenso na escolha de licenças utilizadas ainda, levando-se em conta o fato de todos os repositórios se declararem como de acesso aberto, mesmo contando com dados, tanto abertos quanto restritos e embargados. Vale ressaltar, também, que os *uploads* de dados são todos restritos a instituições membros ou ao registro, com exceção dos *World Clime Exploration and Production Data Bank*, que estão fechados.

3.2 Descrição

A próxima categoria da análise abordou os padrões de metadados para a descrição dos conteúdos disponibilizados nos repositórios de dados de pesquisa estudados.

Metadados – Os padrões de metadados encontrados foram: o Dublin Core, padrão criado para dados bibliográficos; EML para dados ecológicos; a ISO 19115, que orienta dados geográficos; e o DDI, que descreve dados de métodos observacionais de ciências sociais.

Quadro 5 – Descrição: Metadados

Nome do repositório	Padrão de metadados
Base de Dados Científicos da Universidade Federal do Paraná	Dublin Core
CEDAP Research Data Repository – research data	
<i>PPBio Data Repository</i>	<i>EML – Ecological Metadata Language (protocolo NetCDF)</i>
<i>International Ocean Discovery Program</i>	<i>ISO 19115 (formato REST e protocolo OAI-PMH)</i>
<i>IBICT Dataverse Network</i>	<i>DDI – Data Documentation Initiative (formato SWORD)</i>
GLOBE	Outro (próprio)

Fonte: Elaboração própria.

Não foi encontrado, então, qualquer predominância na escolha do padrão de metadados, com inclusive o repositório GLOBE, declarando utilizar um padrão próprio. Não foram encontradas informações a respeito dos metadados, utilizados pelos repositórios *World Clim – Global Climate Data, Exploration and Production Data Bank* e o *Open Research Data @PUC-Rio*.

3.3 Apresentação

Por fim, a categoria “Apresentação” aborda informações sobre os tipos de recursos disponibilizados nos repositórios, como, por exemplo, volume de itens e suas relações.

Tipo de conteúdo – Os tipos de conteúdo encontrados nos repositórios foram: imagens, documentos office padrão, gráficos estruturados, texto sem formatação, texto estruturado, dados científicos ou estatísticos, bases de dados, códigos-fonte, aplicações de *software*, dados arquivados, dados audiovisuais, dados brutos e dados baseados em redes.

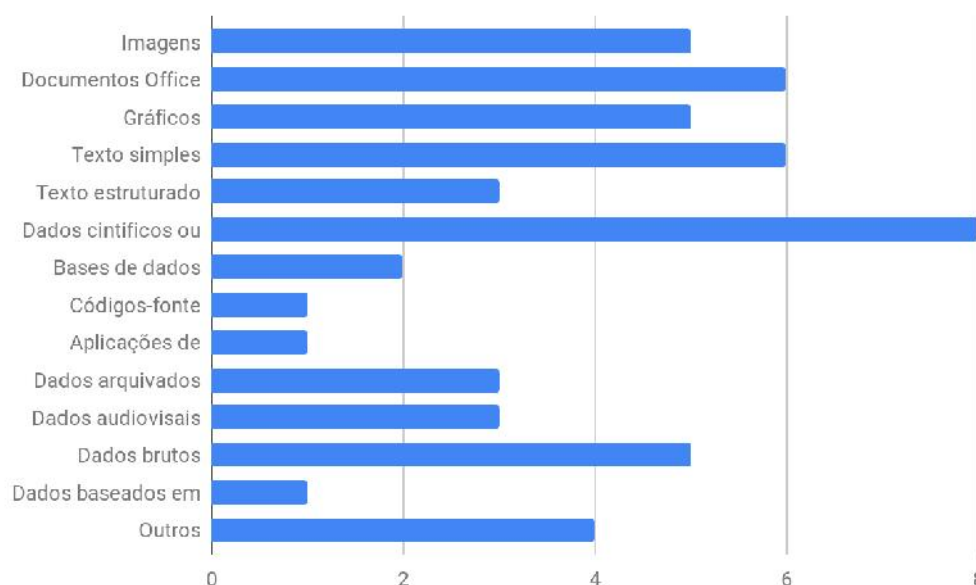
Quadro 6 – Tipos de conteúdo

Nome do Repositório	Tipos de conteúdo
<i>World Clim – Global Climate Data</i>	Gráficos estruturados, dados científicos e estatísticos , outros e dados brutos
<i>International Ocean Discovery Program</i>	Documentos padrão do <i>office</i> , imagens, gráficos estruturados, dados científicos e estatísticos, dados brutos, texto sem formatação e texto estruturado
GLOBE	Documentos padrão do <i>office</i> , dados baseados em rede, gráficos estruturados, dados brutos dados arquivados e outros
<i>IBICT Dataverse Network</i>	Dados científicos e estatísticos, texto sem formatação, outros e dados brutos
<i>PPBio Data Repository</i>	Documentos padrão do <i>office</i> , imagens, dados audiovisuais, texto estruturado, texto sem formatação, dados científicos e estatísticos e dados brutos
<i>CEDAP Research Data Repository – research data</i>	Documentos padrão do <i>office</i> , dados arquivados, texto estruturado, texto sem formatação, dados científicos e estatísticos, dados audiovisuais, imagens e bases de dados
Conclusão	
<i>Base de Dados Científicos da Universidade Federal do Paraná</i>	Texto sem formatação, dados audiovisuais, imagens, dados arquivados, aplicativos de software, gráficos estruturados, código fonte, dados científicos e estatísticos e bases de dados
<i>Exploration and Production Data Bank</i>	Dados científicos e estatísticos, texto sem formatação, gráficos estruturados documentos padrão do <i>office</i> e imagens
<i>Open Research Data @PUC-Rio</i>	Outros dados científicos e estatísticos e documentos padrão do <i>office</i>

Fonte: Elaboração própria.

Podemos, então, perceber uma predominância do tipo: dados científicos ou estatísticos. Os tipos que menos surgem são exatamente as tipificações mais específicas tecnológicas, ou seja, os códigos-fonte, as aplicações de software e os dados baseados em rede:

Imagem 1 – Gráfico de tipos de dados



Fonte: Elaboração própria.

Quantidade de itens – Relativo à quantidade de itens dos repositórios, não foram encontradas informações sobre o *International Ocean Discovery Program*. E sobre o *Exploration and Production Data Bank*, apenas é informado o volume digital, no caso 6 *petabytes*. Para os outros, foi elaborado o quadro a seguir:

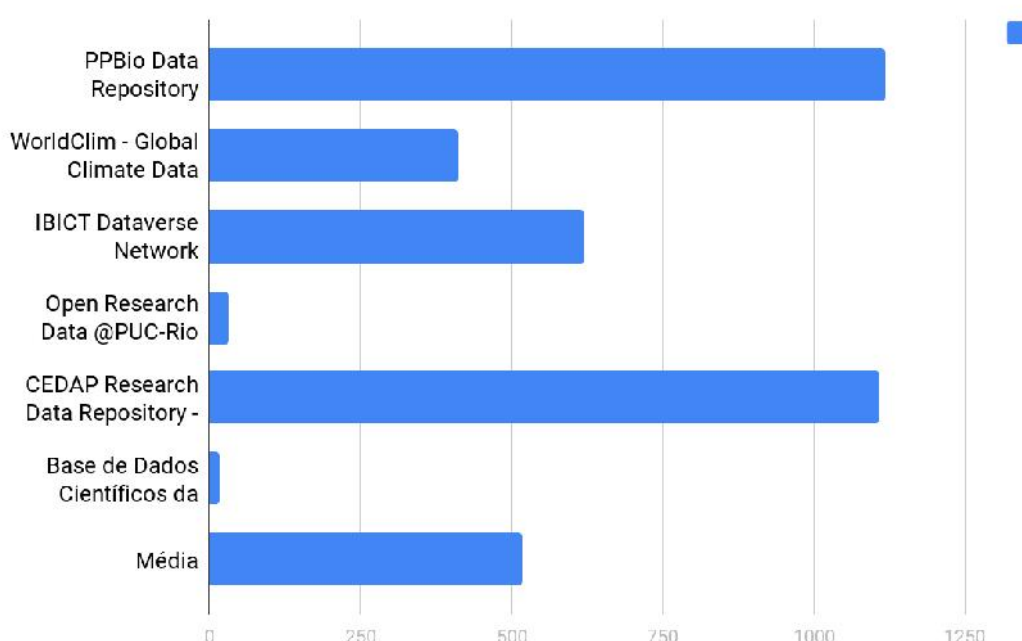
Quadro 7 – Quantidade de itens no repositório.

Nome do repositório	Quantidade de itens
<i>Open Research Data @PUC-Rio-</i>	31
<i>GLOBE</i>	1.444.964
Base de Dados Científicos da Universidade Federal do Paraná	16
<i>PPBio Data Repository</i>	1119
<i>WorldClim</i>	412
<i>IBICT Dataverse Network</i>	619
<i>CEDAP Research Data Repository</i>	1108

Fonte: Elaboração própria.

Para possibilitar a visualização o repositório GLOBE foi excluído do gráfico a seguir, pois sua inclusão impossibilita a visualização da comparação, devido ao seu discrepante número de itens. Podemos perceber que existe uma grande diferença entre o número de itens de cada repositório, mesmo após desconsiderar o repositório GLOBE, com o PPBio apresentando o segundo maior número e a Base de Dados Científicos da Universidade Federal do Paraná, com a menor ocorrência ~~e menor~~. A média de itens seria, então, 515 itens.

Imagem 2 – Gráfico de quantidade de itens.



Fonte: Elaboração própria.

Os problemas de discrepância encontrados podem ter suas origens, exatamente, na própria definição do que é considerado um dado de pesquisa, que ainda não apresenta um consenso amplamente estabelecido.

3.3.3 Relação do item com o conjunto e Relação do conjunto com a publicação

As relações do item podem ser expressas em dois âmbitos, sua relação com o conjunto onde está inserido e sua relação com a publicação, a qual estaria associado. Foram, então, buscadas nos registros encontrados nos repositórios essas ocorrências, resultando na tabela a seguir:

Quadro 8 – Relações

Nome do repositório	Item – Conjunto	Conjunto – Publicação
Exploration and Production Data Bank	Não	Não
Base de Dados Científicos da Universidade Federal do Paraná	Sim	Sim
<i>PPBio Data Repository</i>	Sim	Não
<i>WorldClim</i>	Não	Não
<i>GLOBE</i>	Não	Não
<i>IBICT Dataverse Network</i>	Sim	Sim
<i>International Ocean Discovery Program</i>	Não	Não
<i>Open Research Data @PUC</i>	Sim	Não
<i>CEDAP Research Data Repository</i>	Sim	Não

Fonte: Elaboração própria.

Podemos perceber que cinco, dos nove repositórios, denotam a relação do item com o conjunto, enquanto apenas dois o fazem com a publicação. Totalizando, então, em apenas dois repositórios, que apresentam de forma completa as relações dos itens, enquanto quatro não mostraram relações de forma alguma.

4 CONCLUSÕES

As demandas atuais sobre dados de pesquisa e seu compartilhamento, que advêm da Ciência Aberta, apontam os repositórios de dados, como ferramentas para o armazenamento, organização, compartilhamento e divulgação desses dados. Porém, o baixo número de repositórios, a não padronização e a inconformidade com padrões internacionais estabelecidos são indicativos de que o Brasil tem muito trabalho a ser feito para poder se inserir nas contemporâneas práticas de Ciência Aberta. Outra característica que evidencia a situação é o fato de nenhum dos repositórios analisados possui certificações e apenas metade utiliza identificadores persistentes.

Não há dúvidas de que é necessário realizar um grande trabalho de conscientização da importância da Ciência Aberta para que o País possa contribuir para o avanço célere da Ciência.

REFERÊNCIAS

- COSTA, Michelli Pereira da; LEITE, Fernando César Lima. Repositórios institucionais de acesso aberto à informação científica: proposta de modelo de avaliação. **RECIIS – RevEletron de Comun Infnov Saúde**. Rio de Janeiro, v. 9, n. 3, 9 p. jul./set. 2015.
- GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo : Atlas, 2008.
- RE3DATA, REGISTRY OF RESEARCH DATA REPOSITORIES. **About**. Disponível em: <http://service.re3data.org/about>. Acesso em: 22 de dez. de 2019.
- RODRIGUES, Eloy; *et. al.* **Os repositórios de dados científicos**: estado da arte. 2010. 54 p. Disponível em: <http://repositorium.sdum.uminho.pt/handle/1822/10830>. Acesso em: 22 de dez. de 2019.
- SARMENTO, Fernanda; *et al.* Algumas considerações sobre as principais declarações que suportam o movimento Acesso Livre. **World Congresson Health Information and Libraries**. Salvador: 2005. Disponível em: <<http://hdl.handle.net/10760/8512>>. Acesso em: 22 de dez. de 2019.