cadernos BAD

7ª conferência luso-brasileira acessoaberto Instituto Politécnico de Viseu - Portugal

Repositórios de dados de pesquisa para as ciências da saúde

Lucas Paganine
Universidade de Brasília (UnB)
lnpaganine@hotmail.com

Michelli Costa
Universidade de Caxias do Sul (UCS) e Universidade de Brasília (UnB)
michelli@unb.br

Resumo

O estudo tem por objetivo elaborar um panorama geral dos repositórios de dados de pesquisa na área das ciências da saúde. Para tanto, foi realizada uma pesquisa descritiva, na qual foram analisados vinte repositórios de dados cadastrados e certificados no diretório re3data.org. A análise metodológica foi desenvolvida a partir das características essenciais apontadas por Rodrigues *et al.* (2010) para um repositório de dados, que foram sistematizadas em categorias. As categorias exploram aspectos acerca do armazenamento, descrição e apresentação dos itens. Os resultados do estudo indicam que os repositórios analisados se encontram em diferentes estados de desenvolvimento. No entanto, foi possível observar que não existe tendência para o uso de software de repositório específico, embora haja tendência estruturada para o padrão de metadados e os tipos de recursos disponibilizados. De forma geral, grande parte dos repositórios analisados possuem características de repositórios de acesso aberto, no entanto apenas 30% deles distribuem seus conteúdos acompanhados de licenças abertas. Percebe-se, portanto, uma incipiente e crescente preocupação com o desenvolvimento de sistemas dessa natureza.

Palavras-chave: Repositórios de dados de pesquisa. Ciências da saúde. Open access. Comunicação científica.

Research data repositories for health sciences

Abstract

The study aims to show an overview of research data repositories in the area of health sciences. To achieve it, a descriptive research was held, where twenty registered data and certificated repositories were analysed based in the directory re3data.org. The methodology was based in the essential data repositories characteristics mentioned by Rodrigues *et al.* (2010). The characteristics are related to storage, description and presentation of items. The study`s results indicate that repositories analysed has different stages of development. However, it was observed that there was no trend towards the use of specific repository software, although there is a tendency for the metadata standard and the types of available resources. In general, most repositories have open access characteristics, however only 30% of them distribute their contents based in open license. As conclusion, we can say that the results suggest an incipient and growing concern about the development of such systems.

Keywords: Repositories of research data. Health sciences. Open access. Scientific communication.

Introdução

O advento das novas tecnologias da comunicação e informação provoca mudanças nas formas como as pessoas interagem e se comunicam. Consequentemente, as mudanças têm gerado processos significativos de transformação da comunicação científica. Nesse contexto de mudanças, surge o movimento Open Access, para apresentar uma nova alternativa ao sistema de mercado tradicional dos periódicos científicos comerciais. 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 E SOUZA *et al.*, 2005). A declaração de Bethesada formalizou a demanda para o tratamento dos dados de pesquisa no contexto do acesso aberto.

O compartilhamento de dados de pesquisa é uma questão de crescente importância devido aos seus diversos benefícios para otimização da ciência. Os repositórios de dados de pesquisa têm sido apresentados pela literatura pertinente como uma das ferramentas adequadas para tal objetivo. Porém ainda se apresentam como desafios a padronização dos sistemas e a curadoria dos dados envolvidos. Considerando o contexto brevemente levantado e a importância do tema, o presente estudo buscou identificar a situação atual de desenvolvimento dos repositórios de dados de pesquisa no campo das ciências da saúde.

Metodologia

Para execução da análise proposta, foi realizada uma pesquisa de levantamento e análise descritiva acerca dos repositórios de dados na área das Ciências Saúde, registrados no re3data.org (Registry of Research Data Repositories) e que possuíam algum certificado de sistema. O levantamento dos dados ocorreu no último semestre do ano de 2015.

A partir dos critérios definidos para a pesquisa foram identificados 20 repositórios de dados de pesquisa no campo das Ciências da Saúde (Quadro 1). Dentre os repositórios que compõem o conjunto de análise 50% deles são originários dos Estados Unidos da América, 10% do Canadá e 30% de países da Europa. Apenas dois deles originam-se de países do Sul do mundo, um é de origem indiana e outro de provem de Gana.

QUADRO 1: REPOSITÓRIOS SELECIONADOS PARA ANÁLISE

Nome do repositório	URL
American Type Culture Collection	http://www.lgcstandards-atcc.org/
ArrayExpress	http://www.ebi.ac.uk/arrayexpress/
Bacterial Carbohydrate Structure DataBase	http://csdb.glycoscience.ru/bacterial/index.html
Bii	http://isa-tools.org/
Canadian Epigenetics, Environment and Health Research Consortium Platform	http://www.epigenomes.ca/
Centers for Disease Control and Prevention, Data & Statistics	http://www.cdc.gov/DataStatistics/
ClinicalTrials.gov	http://www.clinicaltrials.gov/
Collaborative Psychiatric Epidemiology Surveys	http://www.icpsr.umich.edu/icpsrweb/CPES/
Danish Data Archive	https://www.sa.dk/en/services/danish-data-archive
DiversityData.org	http://www.diversitydata.org/
Domino	http://mint.bio.uniroma2.it/domino/
Hardin.MD	http://hardinmd.lib.uiowa.edu/
HomoMINT	http://mint.bio.uniroma2.it/HomoMINT/Welcome.do
Human Proteinpedia	http://www.humanproteinpedia.org/
INDEPTH Data Repository	http://www.indepth-ishare.org/index.php/home
InnateDB	http://www.innatedb.com/
MaizeGDB	http://www.maizegdb.org/
Mentha	http://mentha.uniroma2.it/
NeuroMorpho	http://neuromorpho.org/
Neuroscience Information Framework	http://neuinfo.org/

Fonte: Elaboração própria

A análise teve como referência as três características essenciais para um repositório de dados proposta por Rodrigues *et al.* (2010). As características forma utilizadas como categoria de análise dos dados coletados. Portanto, os resultados foram sistematizados em três grandes categorias: armazenamento, descrição e apresentação dos itens (Quadro 2).

QUADRO 2: CATEGORIAS E ITENS ANALISADOS

Categorias	Elementos analisados
Armazenamento	Software: programa(s) utilizados no repositório
	Sistema de preservação: métodos ou práticas de preservação utilizadas
	URL persistente: se o repositório apresenta URL persistente ou não
	Instituição do pesquisador: se o repositório faz referência a instituição à qual o pesquisador pertence ou não
	Licença: tipo de licença(s) utilizada
	Repositório aberto: se o repositório de dados é um repositório de acesso aberto ou não
Descrição	Metadados (padrão): qual padrão de metadados utilizado.
Apresentação dos itens	Tipo de recurso: caracterização dos dados apresentados no repositório, se 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: Elaboração própria

Armazenamento dos dados de pesquisa em repositórios

A primeira categoria de análise tratou de aspectos referentes ao armazenamento dos dados de pesquisa e contemplou questões sobre o *software* dos repositórios, os sistemas de preservação utilizados, as licenças concedidas aos conteúdos armazenados e a política acerca do acesso aberto dos repositórios de dados de pesquisa.

Software dos repositórios de dados de pesquisa nas ciências da saúde

O estudo identificou que existe uma ampla variedade de software utilizado no contexto e que nenhum deles apresenta-se como majoritário em relação aos outros. Dentre os sistemas selecionados para a análise identificou-se que cada repositório apresentou informação singular acerca do software utilizado, conforme indicado no Quadro 3.

QUADRO 3: SOFTWARE UTILIZADOS PELOS REPOSITÓRIOS DE DADOS DE PESQUISA NAS CIÊNCIAS DA SAÚDE

Repositório	Software
HomoMINT	MINT Viewer
Bacterial Carbohydrate Structure DataBase	SOAP
InnateDB	Cerebral program (Java), o ProbeLynx e o INVEX
Mentha	REST
Bii	BioInvIndex
Canadian Epigenetics, Environment and Health Research Consortium Platform	OpenCEMT

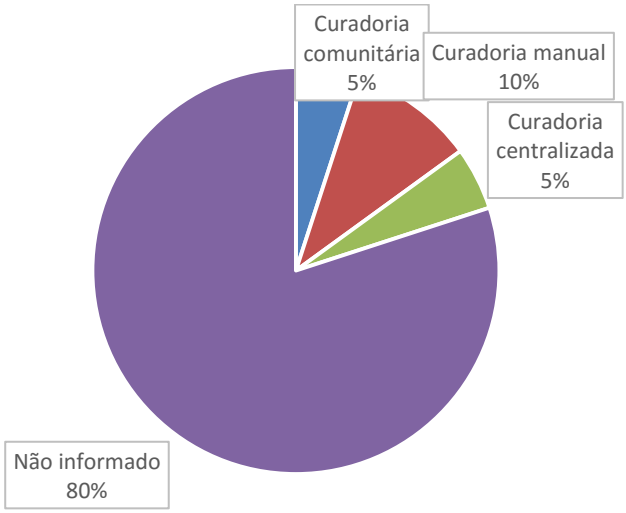
Fonte: Elaboração própria

O HomoMINT com seu software MINT viewer permite uma visualização gráfica no contexto de dados de alto rendimento. Já o software SOAP usado pelo Bacterial Carbohydrate Structure DataBase é uma base de dados relacional baseado no MySQL com uma tabela de aproximação para correlacionar os dados em estruturas. O softaware REST do Mentha é uma interface que oferece uma aplicação gráfica e ferramentas que representam interações e opções de caminhos de navegação. O BioInvIndex do Bii é uma aplicação web, modelo database persistente e pacote de serviços e web serviços. O OpenCent é um software de análise, manuseio e publicação de dados epigenéticos. O repositório InnateDB possui três programas para o desempenho de suas funcionalidades: Cerebral program, ProbeLynx e INVEX. O Cerebra program é um plugin Java para visualização de dados sobre moléculas. O INVEX é utilizado para análise e visualização de parâmetros e metadados dos dados. O ProbeLynx para anotações sobre microarranjos.

Sistemas de preservação dos repositórios de dados de pesquisa nas ciências da saúde

A maior parte dos repositórios analisados (80%) não declararam utilizar sistema de preservação. A falta de informação, no entanto, não permite concluir a ausência de estratégia para tal finalidade, uma vez que essa informação pode apenas ter sido negligenciada no momento do registro do repositório. Dentro os 20% dos repositórios que declararam possuir alguma estratégia de preservação dos conteúdos que armazenam, metade informou que realiza uma curadoria manual, na qual não conta com um sistema de preservação em rede com outras instituições ou iniciativas similares (Figura 1).

FIGURA 1: SISTEMA DE PRESERVAÇÃO DOS REPOSITÓRIOS DE DADOS DE PESQUISA NAS CIÊNCIAS DA SAÚDE



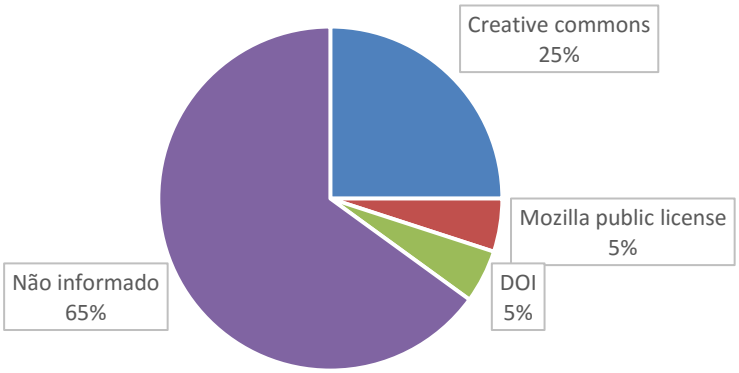
Fonte: Elaboração própria

A ausência de informações acerca do sistema de preservação dos repositórios não é suficiente para afirmar sua inexistência. No entanto, a falta de dados sobre o tema demonstra que ainda não há a preocupação devida com a preservação dos dados de pesquisa.

Licenças de acesso e uso concedidas aos conteúdos armazenados nos repositórios de dados de pesquisa nas ciências da saúde

Apenas seis dos vinte repositórios analisados informaram utilizar alguma licença de acesso e uso para o conteúdo que armazenam e disponibilizam. Dentre o conjunto de licenças previstas pelo re3data.org, a licença Creative Commons foi a mais utilizada, presente em um quarto da totalidade dos repositórios. Menos representativas, as licenças Mozila Public License e a política DOI, foram utilizadas em pelo menos um dos repositórios selecionados (Figura 2).

FIGURA 2: LICENÇAS DE ACESSO E USO CONCEDIDAS AOS CONTEÚDOS ARMAZENADOS



Fonte: Elaboração própria

De forma geral foi observado que apenas 30% dos repositórios de dados de pesquisa nas ciências da saúde utilizam licenças adequadas para o armazenamento, disponibilização e reutilização dos dados no contexto da ciência colaborativa. Frente aos dados obtidos no Diretório é possível perceber que as preocupações acerca do uso de licenças abertas para a distribuição dos conteúdos ainda não alcançaram um nível desejável para os objetivos da ciência aberta.

Política de acesso aberto dos repositórios de dados de pesquisa nas ciências da saúde

Embora tenha sido considerada baixa a utilização das licenças para a ciência aberta, cerca de 85% dos repositórios declararam ser uma iniciativa de acesso aberto. Apenas 3 repositórios, o American Type Culture Collection, Collaborative Psychiatric Epidemiology Surveys e o Danish Data Archive não são se declararam como repositórios de acesso aberto. O American Type Culture Collection disponibiliza tanto os dados quanto o próprio material biológico para compra, o Collaborative Psychiatric Epidemiology Surveys apresenta datasets tanto de uso público quanto de uso restrito e o Danish Data Archive possui alguns dados que só podem ser acessados fisicamente.

Descrição dos dados de pesquisa em repositórios

A segunda categoria de análise abordou os padrões de metadados para a descrição dos conteúdos disponíveis nos repositórios de dados de pesquisa em ciências da saúde.

Metadados dos repositórios de dados de pesquisa nas ciências da saúde

Dentre o conjunto analisado, apenas 8 repositórios apresentaram informações acerca do padrão de metadados utilizado para descrever seus recursos. Em alguns casos notou-se

que um mesmo repositório utiliza mais de um padrão de metadados para a finalidade da descrição. Portanto, foram identificados mais padrões de metadados do que repositórios que apresentavam essa informação. No total foram elencados 9 padrões de metadados, conforme elencado no Quadro 4.

QUADRO 4: PADRÕES DE METADADOS UTILIZADOS PELOS REPOSITÓRIOS DE DADOS DE PESQUISA

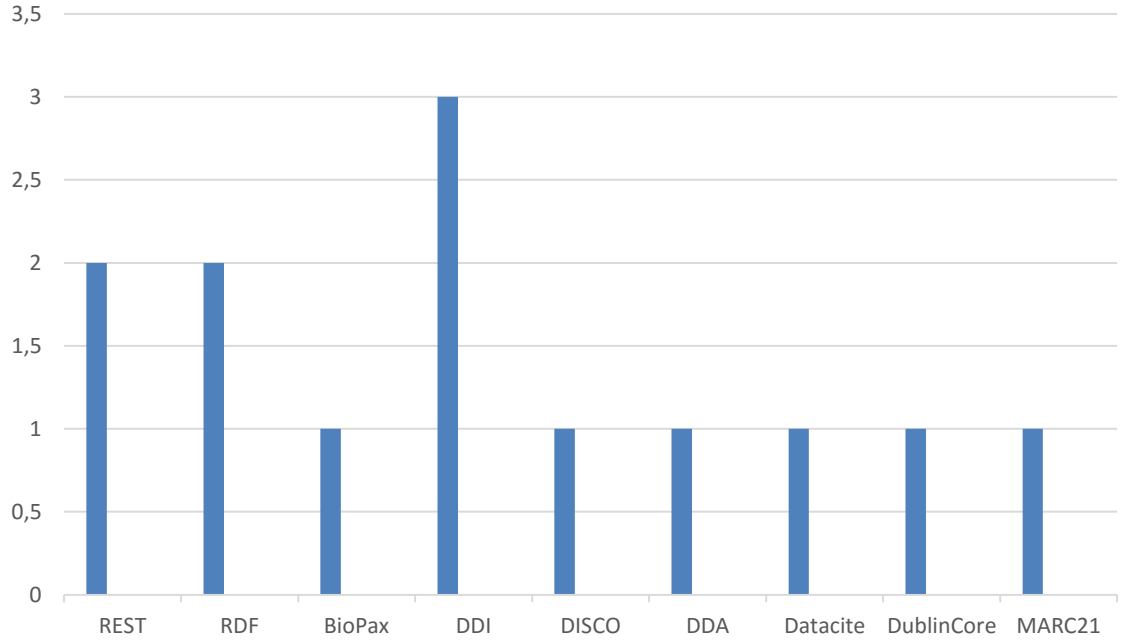
Repositório	Padrão de metadados
ArrayExpress	REST
ClinicalTrials.gov	
Bacterial Carbohydrate Structure DataBase	RDF
INDEPTH Data Repository	DDI, RDF
Danish Data Archive	DDI, DDA, Datacite
Collaborative Psychiatric Epidemiology Surveys	DDI, Dublin Core, MARC21
InnateDB	BioPax
Neuroscience Information Framework	DISCO

Fonte: Elaboração própria

De forma geral, foi identificado que o padrão de metadados mais utilizado pelos repositórios é o Data Documentation Initiative (DDI). O padrão foi desenvolvido pela DDI Alliance para descrição de recursos informacionais relacionados a pesquisas socioeconômicas, censos e outras coleções de microdata. Seu objetivo é ser facilmente compreendido por humanos e computadores, de forma a ter seu uso ampliado (KRAMER; LEAHEY, 2012).

Outros padrões que se desacataram no conjunto analisado foram o RDF e o REST (Figura 1). O RDF é um padrão voltado para facilitar a junção de dados criando modelos simples, já o padrão REST de metadados foca nas interações e restrições destacando dados significantes.

FIGURA 3: FREQUENCIA DE USO DOS PADRÕES DE METADADOS



Fonte: Elaboração própria

Além dos três formatos já descritos, mostraram-se relevantes para o contexto dos repositórios de dados de pesquisa em ciências da saúde os seguintes padrões de metadados: BioPax, DISCO, DDA, Datacite, CublinCore e Marc 21, que serão descritos brevemente. O BioPax faz parte de um projeto aberto e colaborativo e é uma linguagem padronizada que visa possibilitar a interação, troca, visualização e análise de dados biológicos. Já DISCO é um método de integração de informação baseado no protocolo de descoberta da Microsoft para facilitar interoperação entre recursos da internet. O DDA é o padrão próprio do Danish Data Archive. O Datacite é o esquema padrão da organização Datacite para publicação e citação de dados de pesquisa. O DublinCore é um esquema de metadados que objetiva a descrição de objetos digitais a partir de 15 elementos, que podem ser expandidos por seus especificadores. Por fim o MARC21, ou *machine readable cataloging* 21 é uma linguagem que permite a leitura e processamento por máquinas de registros catalográficos, ele se propõe a ser utilizado como formato padrão para troca de registros tanto bibliográficos quanto catalográficos.

Apresentação dos itens nos repositórios

A última categoria de análise sistematizou informações referentes aos tipos de recursos disponíveis nos repositórios, o volume de itens e a relação dos itens como seus conjuntos de dados.

Tipos de recurso disponíveis nos repositórios de dados de pesquisa nas ciências da saúde

De forma geral foram identificados sete tipos de recursos disponíveis nos repositórios analisados. Os tipos em questão correspondem à classificação previamente estabelecida pelo próprio diretório re3data.org. A seguir é apresentada a relação entre os tipos de recursos e os repositórios onde eles estão disponíveis (Quadro 5).

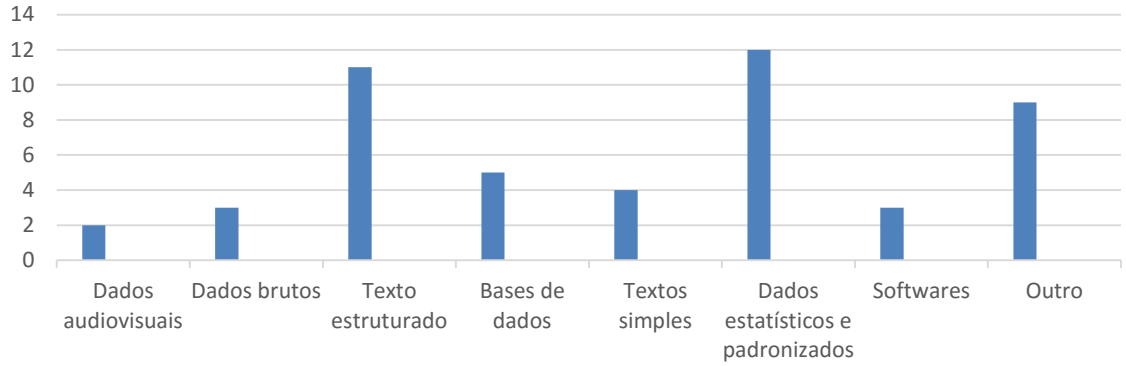
QUADRO 5: TIPOS DE RECURSOS DISPONÍVEIS NOS REPOSITÓRIOS DE DADOS DE PESQUISA EM CIÊNCIAS DA SAÚDE

Tipo de recurso	Repositório de dados de pesquisa
Dados audiovisuais	HomoMINT e NeuroMorpho
Dados brutos	HomoMINT, American Type Culture Collection e DiversityData.org
Texto estruturado	HomoMINT, MaizeGDB, Bacterial Carbohydrate Structure DataBase, Mentha, Human Proteinpedia, Bii, Canadian Epigenetics, Environment and Health Research Consortium Platform, ClinicalTrials.gov, Domino, NeuroMorpho e Neuroscience Information Framework
Bases de dados	American Type Culture Collection, ArrayExpress, INDEPTH Data Repository, Hardin.MD e Neuroscience Information Framework
Textos simples	MaizeGDB, InnateDB, Mentha, e Human Proteinpedia
Dados estatísticos e padronizados	MaizeGDB, ArrayExpress, Bacterial Carbohydrate Structure DataBase, InnateDB, Mentha, INDEPTH Data Repository, Bii, Canadian Epigenetics, Environment and Health Research Consortium Platform, Centers for Disease Control and Prevention, Data & Statistics, Collaborative Psychiatric Epidemiology Surveys, Danish Data Archive e Neuroscience Information Framework
Softwares	InnateDB, Mentha e Neuroscience Information Framework
Outro	MaizeGDB, InnateDB, INDEPTH Data Repository, Human Proteinpedia, Hardin.MD, ClinicalTrials.gov, Collaborative Psychiatric Epidemiology Surveys, DiversityData.org e Domino

Fonte: Elaboração própria

Em relação aos tipos de recursos disponibilizados pelos repositórios foi possível perceber a predominância dos dados estatísticos e padronizados (Figura 4). Este tipo de recurso engloba dados numéricos passíveis ou resultantes de análises estatísticas e dados numéricos estruturados em padrões específicos. Os dados textuais também padronizados em padrões específicos apareceram como o segundo tipo com maior ocorrência dentro do conjunto analisado. Estes resultados apontam que os tipos de recursos pertinentes para as ciências da saúde demandam padrões específicos não previstos pelo diretório re3data.org.

FIGURA 4: FREQUÊNCIA DOS TIPOS DE RECURSOS DISPONIBILIZADOS PELOS REPOSITÓRIOS DE DADOS DE PESQUISA



Fonte: Elaboração própria

Além dos dados estatísticos, padronizados e os textos estruturados, outros tipos de recursos foram identificados com menor relevância para o conjunto de repositórios, são eles: os dados audiovisuais, os dados brutos, as bases de dados, os textos simples e os softwares.

Volume de itens nos repositórios de dados de pesquisa nas ciências da saúde

A união de todos os recursos disponíveis nos repositórios analisados e que possuíam essa informação somam cerca de 840 bilhões de itens. Com relação à quantidade de itens observou-se grande disparidade entre os repositórios, conforme demonstrado no Quadro 6.

QUADRO 6: QUANTIDADE DE ITENS POR REPOSITÓRIO ANALISADO

Nome do repositório	Quantidade de itens
Neuroscience Information Framework	829.679.866
Human Proteinpedia	7.038.972
ArrayExpress	1.962.932
InnateDB	888.753
Mentha	700.745
HomoMINT	343.277
ClinicalTrials.gov	206.902

American Type Culture Collection	74.000
Bacterial Carbohydrate Structure DataBase	21.904
Domino	15.981
Collaborative Psychiatric Epidemiology Surveys	6.274
INDEPTH Data Repository	4.239
Bii	133

Fonte: Elaboração própria

Os repositórios MaizeGDB, Hardin.MD, Canadian Epigenetics, Environment and Health Research Consortium Platform, Centers for Disease Control and Prevention, Data & Statistics, Danish Data Archive, DiversityData.org e o NeuroMorpho não informam o número de itens em seus bancos de dados.

A distância tão discrepante no número de itens entre os repositórios deve-se, em parte, ao escopo e as políticas adotadas. Enquanto alguns consideram um conjunto de dados completo como um único item, outros tratam cada molécula e cada interação como um item diferente, como nos casos dos Neuroscience Information Framework e da Human Proteinpedia. Considerando as peculiaridades ressalvadas, o Neuroscience Information Framework, é responsável por 98% do universo, mostrando-se assim como o repositório mais expressivo do universo analisado.

Relação dos dados com seu conjunto

A maioria dos repositórios (70%) apresentam informações sobre a relação do item com o seu conjunto (dataset). De acordo com Padilla Navarro *et al.* (2013) o conjunto de dados é uma representação sistemática parcial do objeto que está sendo pesquisado. Além disto, Rodrigues *et al.* (2010) ressaltam que eles são caracterizados por reunir informações ou fatos relacionados entre si e registrados num formato comum. Portanto, a existência de conjunto de dados no repositório demanda que a visão sobre objetos individuais seja extrapolada e que o conjunto receba tratamento diferenciado em casos específicos (RODRIGUES et. al., 2010).

Foram sete os repositórios que não apresentam informações sobre a relação dos itens com suas coleções, a saber: American Type Culture Collection, Mentha, Hardin.MD, Bii, Canadian Epigenetics, Environment and Health Research Consortium Platform e o Domino. Destaca-se que nos casos onde não foi possível identificar essa relação não é possível concluir que a relação é inexistente, ela apenas pode não ter sido devidamente informada no registro do repositório.

Considerações finais

Os resultados encontrados demonstram que os repositórios de dados de pesquisa na área das ciências da saúde encontram-se em diversos estados de desenvolvimento. De um lado, foram observados repositórios em estados avançados como o Bacterial Carbohydrate Strcture DataBase, que apresentou interface bem trabalhada e estruturação na descrição e apresentação de seus itens. De outro lado, foram identificados repositórios menos desenvolvidos, como o Danish Data Archive, que apresenta problemas sérios de navegação, impedindo muitas vezes a coleta de dados para análise proposta. No entanto, o volume dos itens nos repositórios indica uma crescente preocupação com a disponibilização dos dados de pesquisa em repositórios.

Ressalta-se ainda que a pesquisa se limitou aos repositórios de dados de pesquisa classificados como da área das ciências da saúde, que possuíam algum certificado de sistema e que estavam cadastrados no diretório re3data.org. Além disto, as fontes de informações consideradas para a análise limitaram-se às informações disponibilizadas pelos próprios portais dos repositórios e aquelas disponíveis no diretório re3data.org. Apesar das limitações declaradas sobre a pesquisa, considera-se que o retrato apresentado sobre o cenário tem condições de instrumentalizar as discussões sobre o compartilhamento e o reuso dos dados de pesquisa para a promoção da ciência aberta e produzir orientações para o aprimoramento dos sistemas, em especial na área das ciências da saúde.

Referências bibliográficas

DDI ALLIANCE. Document, discover, and interoperate. Disponível em: <http://www.ddialliance.org/>. Acesso em: 25 de abr. de 2016.

DECLARAÇÃO DE BETHESDA. Meeting on Open Access Publishing, Bethesda. Abril. 2003. Disponível em: <http://legacy.earlham.edu/~peters/fos/bethesda.htm>. Acesso em: 03 de fev. de 2016.

KRAMER, S.; LEAHEY, A. Using RDF to describe and link social science data to related resources on the Web: leveraging the Data Documentation Initiative (DDI) model - Parade@Portsmouth. . In: DATA DOCUMENTATION INITIATIVE. Ann Arbor, Michigan: 2012Disponível em: <http://eprints.port.ac.uk/9029/>. Acesso em: 12 maio. 2016

OECD. OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007. Disponível em: <http://www.oecd.org/dataoecd/9/61/38500813.pdf>

PADILLA NAVARRO, P. A. *et al.* Acceso a datos de investigación e información científica en Chile. Revista española de Documentación Científica, v. 36, n. 3, 2013.

PANTON PRINCIPLES, Principles for Open Data in Science. About. Disponível em: <http://pantonprinciples.org/>. Acesso em: 22 de dez. de 2015.

REGISTRY OF RESEARCH DATA REPOSITORIES, RE3DATA. About. Disponível em: <http://service.re3data.org/about>. Acesso em: 22 de dez. de 2015.

RODRIGUES, E. *et al.* Os repositórios de dados científicos : estado da arte. [s.l: s.n.]. Disponível em: <http://repositorium.sdum.uminho.pt/handle/1822/10830>. Acesso em: 17 abr. 2015.

SARMENTO E SOUZA, M. F. *et al.* Algumas considerações sobre as principais declarações que suportam o movimento Acesso Livre. 2005Disponível em: <http://eprints.rclis.org/8512/>. Acesso em: 30 mar. 2016

W3C. Resource Description Framework (RDF). Disponível em: <https://www.w3.org/RDF/>. Acesso em: 25 de abr. de 2016.