



Universidade de Brasília – UnB  
Faculdade de Ciência da Informação – FCI

# **AVALIAÇÃO DE REPOSITÓRIOS DE DADOS DE PESQUISA NA ÁREA DE CIÊNCIAS DA SAÚDE**

Lucas Nóbrega Paganine

Orientadora: Michelli Pereira da Costa

Brasília  
Junho de 2016

Lucas Nóbrega Paganine

**AVALIAÇÃO DE REPOSITÓRIOS DE DADOS DE PESQUISA  
NA ÁREA DE CIÊNCIAS DA SAÚDE**

Monografia apresentada à banca examinadora como requisito parcial para a conclusão do curso de Biblioteconomia da Faculdade de Ciência da Informação, Universidade de Brasília.

Orientadora: Michelli Pereira da Costa

Brasília  
Junho de 2016



**Título: Avaliação de repositório de dados de pesquisa na área de Ciências da Saúde.**

**Aluno: Lucas Nóbrega Paganine.**

Monografia apresentada à Faculdade de Ciência da Informação da Universidade de Brasília, como parte dos requisitos para obtenção do grau de Bacharel em Biblioteconomia.

Brasília, 08 de julho de 2016.

**Michelli Pereira da Costa - Orientadora**

Professora da Faculdade de Ciência da Informação (UnB)  
Mestre em Ciências da Comunicação

**Fernanda Passini Moreno - Membro**

Professora da Faculdade de Ciência da Informação (UnB)  
Doutora em Ciência da Informação

**Fernando César Lima Leite - Membro**

Professor da Faculdade de Ciência da Informação (UnB)  
Doutor em Ciência da Informação

Dedico este trabalho a meu pai, que mesmo não estando mais ao nosso lado sempre estará presente guiando cada um dos meus passos.

## **Agradecimentos**

À minha família, por sempre estarem presentes e me apoiar. Às minhas tão amadas mãe e irmã, não há palavras o suficiente para expressar o quanto merecem agradecimentos, mas posso dizer que sei que sem o carinho e preocupação delas não estaria nem perto de onde estou hoje. Às minhas queridas tias, Rosa, Carolina, Joseana, Luzia e minha madrinha Patrícia que além de sempre estarem dispostas a ajudar, foram, sendo exemplos a seguir, a minha maior fonte de inspiração.

Aos meus amigos, que sendo muitos para citar aqui, sabem de quem estou falando, que nunca falham em me mostrar que em qualquer situação há algo de bom a se ganhar. Em geral a todos que me cercam e amam e puxaram a minha orelha quando eu precisei.

Em especial a minha orientadora Michelli, não poderia ter sonhado com alguém melhor para me acompanhar nessa etapa final, com sua empolgação contagiante deixa impossível não se apaixonar pelo tema, e que com sua paciência e compreensão torna tudo mais agradável.

*“Um coração de aço conquista o mundo.”*

(Desconhecido)

PAGANINE, Lucas Nóbrega. **Avaliação de repositórios de dados de pesquisa na área de ciências da saúde**. 2016. 78 p. il. Monografia (Bacharelado em Biblioteconomia) — Brasília: Universidade de Brasília, 2016.

## RESUMO

Esse trabalho, avaliação de repositórios de dados de pesquisa na área de ciências da saúde, é um estudo que tem por objetivo elaborar um panorama geral dos repositórios de dados de pesquisa na área das ciências da saúde. Para tal, foi realizada uma pesquisa descritiva, na qual foram analisados vinte repositórios de dados certificados cadastrados no RE3DATA. A análise foi desenvolvida a partir das três características essenciais para um repositório de dados encontradas no documento propostos por Rodrigues e Saraiva (2010). Como parâmetro de comparação, foi utilizado o documento *Principles and guidelines for access to research data from public funding* publicado pela OECD em 2007. A partir das características principais, foram definidas 3 grandes categorias que por sua vez foram divididas em subcategorias. As categorias exploram aspectos acerca do armazenamento, descrição e apresentação dos itens. Os resultados indicam que esses repositórios de uma forma geral se encontram em diferentes estados de desenvolvimento. Porém já se percebe uma crescente preocupação com a disponibilização dos dados de pesquisa devido a avanços em diversos fatores, demonstrando que futuramente esses dados terão cada vez mais visibilidade e uma maior demanda, trazendo diversos benefícios ao processo científico.

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

## **ABSTRACT**

This work, evaluation of research data repositories in the field of health sciences, is a study that aims to develop an overview of research data repositories in the area of health sciences. To this end, a descriptive research was carried out, in which were analyzed twenty certified data repositories registered in RE3DATA. The methodological analysis was developed from the three essential characteristics for a data repository found in the document proposed by Rodrigues and Saraiva (2010). As a benchmark, we used the document Principles and guidelines for access to research data from public funding published by the OECD in 2007. From these main features 3 major categories were defined, which in turn were divided into subcategories. Categories explore aspects concerning the storage, description and presentation of items. The results indicate that these repositories in general are at different stages of development. But we already see a growing concern about the availability of research data due to advances on several factors, demonstrating that in the future this data will have increased visibility and greater demand, bringing many benefits to the scientific process.

**Keywords:** Research data repositories. Health Sciences. Open access. Scientific communication.

## LISTA DE FIGURAS

Figura 1- Scientific Communication: A Model for 2020 ..... p. 17

## LISTA DE TABELAS

**Tabela 1-** Comparação entre os textos das declarações de Budapeste, Bethesda e Berlim que regem o Acesso Livre ao conhecimento ..... p. 21

**Tabela 2-** Distribuição de artigos de acesso aberto (OA) por país do primeiro autor ..... p. 29

**Tabela 3-** Identificação ..... p. 33

**Tabela 4-** Armazenamento ..... p. 34

**Tabela 5-** Descrição ..... p. 34

**Tabela 6-** Apresentação dos itens ..... p. 34

## LISTA DE QUADROS

**Quadro 1-** Princípios OECD para acesso a dados de pesquisa .....p. 23

**Quadro 2-** Quadro da metodologia ..... p. 35

**Quadro 3-** Software ..... p. 42

**Quadro 4-** Sistemas de preservação ..... p. 43

**Quadro 5-** Padrão de metadados ..... p. 47

**Quadro 6-** Pontos fortes ..... p. 54

**Quadro 7-** Dimensões de análise: Identificação ..... p. 60

**Quadro 8-** Dimensões de análise: Armazenamento ..... p. 61

**Quadro 9-** Dimensões de análise: Descrição ..... p. 62

**Quadro 10-** Dimensões de análise: Apresentação do item ..... p. 63

## LISTA DE GRÁFICOS

**Gráfico 1-** Disponibilidade do texto dos artigos selecionados (n=4,667) ..... p. 30

**Gráfico 2-** Tipo ..... p. 40

**Gráfico 3-** Países ..... p. 41

**Gráfico 4-** Sistema de preservação ..... p. 44

**Gráfico 5-** Instituição do pesquisador ..... p. 45

**Gráfico 6-** Licenças ..... p. 46

**Gráfico 7-** Acesso aberto ..... p. 47

**Gráfico 8-** Padrão de metadados ..... p. 48

**Gráfico 9-** Tipo de recurso ..... p. 51

**Gráfico 10-** Relação do item com o conjunto ..... p. 53

**Gráfico 11-** Relação do conjunto com a publicação ..... p. 54

## **LISTA DE SIGLAS**

AA- Acesso Aberto

OA- Open Access

OAI- Open Archives Initiative

BIREME- Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde Área de Gestão do Conhecimento, Bioética e Pesquisa

BOAI- Budapest Open Initiative

BVS- Biblioteca virtual em saúde

CNPq- Conselho nacional de desenvolvimento científico e tecnológico

EUA- Estados Unidos da América

FAPESP- Fundação de amparo à pesquisa do estado de São Paulo

HINARI- Health InterNetwork Access to Research Initiative

NIH- National Institutes of Health

OECD- Organisation for Economic Co-operation and Development

OMS- Organização Mundial da Saúde

OPAS- Organização Pan-Americana da Saúde

OSI- Open Society Institute

RE3DATA- Research Data Repositories

RRI- Research Related Information

SCIELO- Scientific Electronic Library Online

TIC- Tecnologias da Informação e Comunicação

## SUMÁRIO

<b>1. INTRODUÇÃO .....</b>	<b>12</b>
1.1. OBJETIVOS .....	13
1.2. JUSTIFICATIVA .....	13
<b>2. REVISÃO DE LITERATURA.....</b>	<b>15</b>
2.1. COMUNICAÇÃO CIENTÍFICA .....	15
2.2. MOVIMENTO OPEN ACCESS .....	18
2.3. COMPARTILHAMENTO DE DADOS DE PESQUISA E OS REPOSITÓRIOS DE DADOS.....	22
2.4 CIÊNCIAS DA SAÚDE NO OPEN ACCESS.....	25
<b>3. METODOLOGIA .....</b>	<b>32</b>
3.1. UNIVERSO E AMOSTRA .....	32
3.2. FONTES DE INFORMAÇÃO E TÉCNICA DE COLETA DE DADOS.....	32
3.3. TÉCNICA DE ANÁLISE DOS RESULTADOS.....	33
<b>4. DISCUSSÃO DOS RESULTADOS .....</b>	<b>36</b>
4.1. DESCRIÇÃO E ANÁLISE .....	36
4.1.1. IDENTIFICAÇÃO .....	36
4.1.2. ARMAZENAMENTO.....	41
4.1.3. PRESERVAÇÃO .....	42
4.1.4. REFERÊNCIA A INSTITUIÇÃO DO PESQUISADOR.....	44
4.1.5. LICENÇAS DE ACESSO E USO .....	45
4.1.6. INICIATIVA DE ACESSO ABERTO .....	46
4.1.7. PADRÃO DE METADADOS .....	47
4.1.8. DESCRIÇÃO DOS RECURSOS.....	49
4.1.9. TIPOS DE RECURSO.....	50
4.1.10. TAMANHO DOS REPOSITÓRIOS .....	51
4.1.11. RELAÇÃO DOS DADOS COM SEU CONJUNTO .....	52
4.1.12. RELAÇÃO DO CONJUNTO DE DADOS DA PUBLICAÇÃO .....	53
<b>5. CONCLUSÕES.....</b>	<b>55</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS.....</b>	<b>57</b>
<b>APÊNDICE A.....</b>	<b>60</b>
<b>APÊNDICE B.....</b>	<b>61</b>
<b>APÊNDICE C.....</b>	<b>62</b>
<b>APÊNDICE D.....</b>	<b>63</b>

## 1. Introdução

Com a evolução da comunicação científica, diversas mudanças, advindas do avanço tecnológico, ocorreram gerando uma intensificação do acesso e partilha de informações. Atualmente o paradigma atual da comunicação científica apresenta a questão da necessidade de uma nova forma de tratamento do processo como um todo, com novas ferramentas e modelos, demandando novos métodos de gestão e estratégias.

Nesse contexto de mudanças, surge o movimento Open Access, para apresentar uma nova alternativa ao sistema de mercado tradicional dos periódicos científicos comerciais. O movimento se opõe às restrições ao acesso às publicações científicas. Foram três os eventos mais significativos para o Open Access, resultando nas declarações de Budapeste (2002), Bethesda (2003) e por fim Berlim (2003). Das três declarações, a de Bethesda recebe foco especial pois trata também de dados de pesquisa e resultou de uma reunião de profissionais ligados à área das ciências da saúde.

As ciências da saúde sempre foram vanguardistas ao se tratar de sistemas de informação, e com relação ao Open Access não foi diferente, diversas iniciativas como por exemplo o Scielo (que hoje abrange outras áreas de pesquisa) e PubMed Central são voltadas ao 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 tem sido apresentados pela literatura como as ferramentas adequadas para tal objetivo. Porém ainda se destacam como desafios para o contexto a padronização dos sistemas e a curadoria dos dados envolvidos. O que nos leva então à seguinte pergunta: *Qual panorama em que esses repositórios de dados de pesquisa, mais especificamente para a ciência da saúde, se encontram?*

O trabalho está organizado em 5 grandes seções e seus respectivos subtópicos, são eles: 1- a Introdução, subdividida em Objetivos e Justificativa, 2- a Revisão de literatura, com os 4 temas, Comunicação científica, Movimento OA, Compartilhamentos de dados de pesquisa e os repositórios de dados e Ciências da saúde no OA. A terceira seção é 3- Metodologia, dividida em Universo e amostra, Fontes de informação e técnica de coleta de dados e Técnica de análise dos resultados, 4- Discussão dos resultados, dividida em Lista de repositórios e Descrição e análise, que por sua vez foi subdividida nos quesitos estabelecidos e por fim a quinta seção Conclusões.

### **1.1. Objetivos**

O objetivo geral deste trabalho é traçar um panorama dos repositórios de dados de pesquisa na área de ciências da saúde. Para tratar da questão foram estabelecidos 3 objetivos específicos, são eles:

1. Levantar os repositórios existentes na área de ciências da saúde.
2. Descrever os repositórios selecionados.
3. Analisar de forma comparativa os repositórios descritos.

### **1.2. Justificativa**

A comunicação científica é um processo essencial dentro do desenvolvimento da ciência e sua evolução melhora a prática da ciência de forma geral (MEADOWS, 1999). Atualmente essa comunicação vem sofrendo transformações impulsionadas principalmente pelo avanço tecnológico, que acabaram resultando em uma mudança de perspectiva da própria comunicação científica. Essa nova perspectiva valoriza a velocidade e abrangência da troca de informações e dados, proporcionando maior alcance da disseminação de pesquisas e descobertas, diminuindo assim as chances de repetição de trabalhos e auxiliando na coleta de informações e dados.

Nesse novo paradigma de comunicação científica, os dados de pesquisa merecem uma atenção especial, pois uma partilha eficiente desses dados colabora para alcançar as novas metas da comunicação científica. Propiciando benefícios como a otimização de recursos científicos, de forma a evitar a duplicação de esforço.

As estruturas vigentes que suportam e dão apoio a esses dados são atualmente os repositórios de dados de pesquisa. Repositórios são sistemas de informação que permitem a coleta, organização, armazenamento, preservação e a recuperação de dados brutos de pesquisa e fornecem informações relevantes sobre esses dados, como por exemplo método de coleta e significância do dado com relação ao conjunto ou à publicação. Porém estes repositórios, devido a particularidades de cada área do conhecimento e das diversas iniciativas existentes ou da falta do devido destaque à questão, não se encontram padronizados.

A diversidade de sistemas de informação desta natureza faz com que surjam diferentes modelos demandando então estudos para avalia-los de maneira a aferir suas qualidades, problemas, pontos fortes e fracos e outros aspectos. Logo, os estudos tem potencial para contribuir para desenvolver referências que futuramente podem ajudar a proporcionar

melhorias e resolver problemas relacionados a estes sistemas e os contextos em que estão inseridos.

A área da ciência da saúde foi escolhida por ser proeminente em questões de inovações em sistemas de informação. Especialmente considerando-se o tema tratado, como exemplificado pela ocorrência da reunião de Bethesda, onde profissionais da área biomédica já discutiam o acesso livre em 2003, destacando os dados de pesquisa e a necessidade de sua distribuição juntamente com as pesquisas.

O tema foi escolhido em conjunto com a orientadora desse trabalho, devido à atual crescente demanda sobre o assunto e destaque recebido atualmente, ele partiu de uma experiência de pesquisa (PAC) onde a própria orientadora participou e ajudou a organizar. Na próxima seção encontraremos a Revisão de literatura abordando um pouco da Comunicação científica, do movimento OA, do compartilhamento de dados de pesquisa e os repositórios de dados e por fim as Ciências da saúde no OA.

## 2. Revisão de literatura

### 2.1. Comunicação científica

Comunicação científica, de acordo com Garvey (2014, p. IX)<sup>1</sup> trata das atividades de troca de informação que ocorrem principalmente entre cientistas envolvidos ativamente na frente de pesquisa. Ela se faz especialmente presente durante processo de investigação e durante a divulgação dos resultados. Podendo ocorrer principalmente por meio de periódicos científicos, mas também por documentários, plataformas *online* como *blogs* ou correios eletrônicos, entre outros. Roosendaal e Geurts (1998, p.6) defendem que ela serve para impulsionar o progresso da pesquisa objetivando o crescimento do conhecimento de uma forma geral aumentando a efetividade e eficiência dessa mesma pesquisa.

Medeiros e Caregnato (2012, p. 313) afirmam que a ciência universal precisa de pesquisadores de diversas áreas do conhecimento trabalhando em conjunto. Consideram ainda que esse processo de colaboração se dá partindo de um outro processo social entre indivíduos engajados em algum trabalho, compartilhando e cooperando para a geração de conhecimento. Ainda de acordo com os autores, esse processo todo acaba por desenvolver outros cada vez mais complexos. Esse processo de colaboração teria recentemente se intensificado devido a transformações na própria ciência e na facilidade de acesso à comunicação científica, especialmente devido às tecnologias de informação e comunicação.

Novas tecnologias de comunicação, segundo Costa (2014, p. 43) propiciam uma maior integração informal de pesquisadores, maior acesso ao material bibliográfico e criar melhores condições de visibilidade aos trabalhos dos pesquisadores. A autora ainda afirma, que transformações notáveis ocorreram nos processos de comunicação científica. Especialmente com relação às atividades de disseminação de resultados de pesquisas e na celeridade do processo de forma geral, referindo-se ao movimento open access (OA).

Medeiros e Caregnato (2012, p. 312) também comentam no mesmo sentido, porém de uma forma mais generalista ao dizer “Os serviços disponibilizados em rede trouxeram grande suporte para o intercâmbio de dados e informações entre pesquisadores e instituições.”.

Costa (2014, p. 15) explica o desenvolvimento desse progresso a partir da crise dos periódicos somado ao contexto das novas TICs. Quando esses canais eletrônicos acabaram por ganhar mais espaço dentro da comunicação científica, com novas tecnologias promovendo o nascimento de periódicos científicos eletrônicos e bibliotecas digitais com acesso aberto. Embora a soma das TICs e as reivindicações do movimento OA tenham sido

---

<sup>1</sup> Edição digital do livro, disponível no Google Books

significantes para a modernização da comunicação científica Medeiros e Caregnato (2012, p. 315) afirmam que

não necessariamente foi o avanço tecnológico que institui o compartilhamento e exploração de dados, mas sim os impulsionou, uma vez que esses procedimentos já eram realizados quando não se dispunha de infraestruturas para abrigar grandes bases de dados, como, por exemplo, através da comunicação informal entre pares

A possibilidade de compartilhamento e exploração de dados já existiam na comunicação científica. No entanto, o impulso das TICs possibilitou um maior alcance de comunicação e interação, como explicitado por Medeiros e Caregnato (2012, p. 320) “A perspectiva de compartilhar dados entre cientistas que estão distribuídos por todo o mundo pode fazer com que as interações e possibilidades de colaboração se tornem cada vez mais reais”.

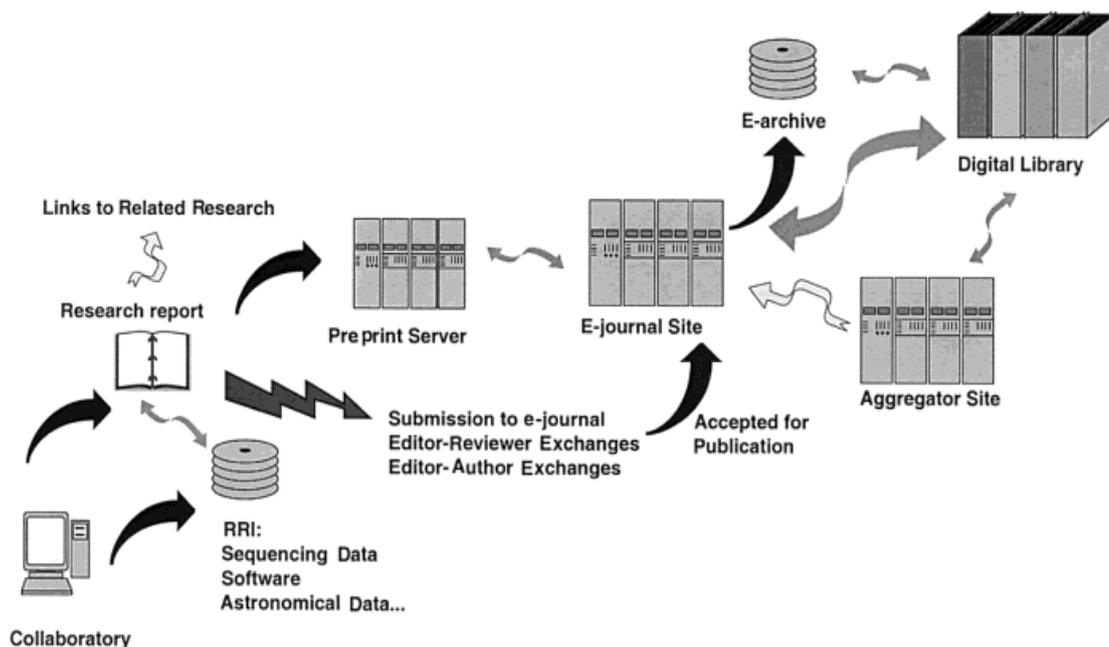
Um exemplo dessas mudanças apresentado por Roosendaal e Geurts (1998, p. 7) é o fato da informação não ser mais escassa. Ao invés disso o elemento escasso atualmente é o tempo e recursos gastos na pesquisa ao explorar a diversidade de informações. Eles ainda frisam que o valor real não está mais na informação em si mesma mas sim na sua comunicação efetiva e eficiente e no gerenciamento do conhecimento. Isso faz com que a pesquisa estratégica seja inerente ao conceito de comunicação científica e uma questão fundamental.

Dentro desse contexto de mudanças paradigmáticas Roosendaal e Geurts (1998, p. 2) ainda ressaltam a necessidade de novos métodos de comunicação científica. Pois elas levariam em consideração o tempo de pesquisa necessário para se explorar a quantidade de informações como fator chave na eficiência e eficácia do sistema como um todo e acelerariam o avanço científico. Para Roosendaal e Geurts (1998, p. 2) essa mudança essencial seria estratégica e não tática, proporcionando muito mais do que um mero melhoramento do sistema da cadeia de informações.

Grande parte dessa transformação estaria englobada dentro dos conceitos de curadoria da informação, que estão intimamente ligados à essa situação de gerenciamento de dados científicos. Pois a curadoria possibilita uma agregação de valores à dados de pesquisa, além de seu compartilhamento e reuso. As práticas consideram ainda a preservação e manutenção de registros (MEDEIROS; CAREGNATO, 2012 p. 319).

Diversas ferramentas para tentar suprir essa demanda surgiram, tal como exemplifica o modelo de comunicação científica proposto por Hurd em 2000 (Figura 1). Baseado no de Garvey e Griffith de 1979, foram destacados 3 grandes procedimentos: produção, disseminação e uso. Todos eles levam em consideração as mudanças econômicas, sociais e políticas ocasionadas pela transição do impresso para o eletrônico (COSTA, 2014 p. 41).

**Figura 1-** Scientific Communication: A Model for 2020



Fonte: Hurd (2000, p. 1281)

Esse modelo ficou conhecido como “um modelo para 2020” e abrange aspectos advindos do avanço tecnológico informacional que o de Garvey e Griffith não abordaram. O modelo inclui por exemplo bases de dados bibliográficos online e textos passíveis de leitura por máquina (HURD, 2000, p. 1280). Nele também é defendido a consideração da redefinição dos papéis de organizações na tentativa de se adaptar ao paradigma vigente e do surgimento dos chamados serviços agregadores, que, baseando-se em assinaturas eletrônicas, oferecem um mecanismo de busca comum para um grande número de periódicos.

Dentre outros elementos, o modelo aborda os dados de pesquisa, utilizando a terminologia de RRI- research related information (informação relacionada à pesquisa). Considera-se que os dados brutos, antes obtidos por telefones, faxes e visitas físicas agora se encontram em servidores acessíveis online para utilização, inclusive à muitos que antes anteriormente não teriam acesso a eles. (HURD, 2000, p. 1281)

Surgiram também infraestruturas que permitem o acesso, compartilhamento e reuso remoto a dados científicos primários ou dados brutos (MEDEIROS; CAREGNATO, 2012, p. 315). Dentre essas diversas inovações o movimento open access (OA) se destaca pela sua

filosofia de compartilhamento e acesso, que se estende também aos dados de pesquisa, podendo nesse caso ser chamado de movimento open data.

## 2.2. Movimento Open Access

Costa e Leite (2015, p. 3) destacam que a natureza das articulações desse movimento internacional se baseiam em ações que gerem resultados, disponibilizando e permitindo o acesso à produção científica em acesso aberto do mundo inteiro. O objetivo das iniciativas “representa a disponibilização livre e irrestrita das publicações científicas, em texto completo, por meio da Internet. A expressão open access foi utilizada pela primeira vez para este propósito pela BOAI” (COSTA, 2014, p. 43).

O movimento de acesso aberto surge como “uma resposta da comunidade científica que até então era “refém” do sistema de mercado estabelecido pelas editoras científicas” (FERREIRA, 2012, p. 32). Baptista *et al.* (2007, p.2) destacam como benefício de tal movimento “a maior visibilidade das pesquisas e sua utilização pelo maior número possível de interessados”

O OA é definido por Costa (2014, p. 48) como “disponibilização das publicações científicas de modo online, livre e sem restrições.” e por Rodrigues *et al.* (2010, p. 47) como “acesso à literatura científica em formato eletrônico, em particular aos artigos de revistas com revisão por pares, sem barreiras de preço e de permissões.”. Brody *et al.* (2007, p. 1) ressaltam ainda que OA é sinônimo de acesso gratuito e online a publicações, podendo se estender para os dados produzidos ou recolhidos. Nos três casos percebe-se que o acesso aberto se trata da facilitação de acesso, abrangendo custos, formato (geralmente eletrônico e online) e a eliminação de outras barreiras.

Por mais que o termo só tenha sido cunhado na Budapest Open Access Initiative (BOAI) seus conceitos já permeavam a comunidade científica. Um exemplo de acordo com Sarmiento *et al.* (2005, p. 3) é a Declaração de Santo Domingo em 1999. Resultante da Primeira Reunião Regional de Consulta da América Latina e do Caribe, relacionada à Conferência Mundial sobre a Ciência para o séc. XXI: uma visão nova e uma base de ação, ela já abordava tais ideais dentro do tópico “ciência para todos”. No mesmo ano, a ideia se consolidou com a criação da Open Archives Initiative (OAI) em 1999 a partir de uma reunião, a Convenção de Santa-Fé, com responsáveis por repositórios de e-prints acadêmicos que já objetivava transformar a comunicação científica (FERREIRA, 2012, p. 31).

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 Costa e Leite (2015, p. 3) os descrevem em linhas gerais dizendo 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.

A primeira declaração, a BOAI, de acordo com Sarmiento *et al.* (2005, p. 3 e 4) foi resultante de uma reunião da Open Society Institute (OSI) em 2001 e

teve como objetivo discutir a questão do acesso livre a literatura científica; analisar como as iniciativas isoladas poderiam trabalhar em conjunto de forma a atingir um sucesso mais amplo, profundo e rápido, e verificar como a OSI (instituto que pertence a Soros Foundation) e outras fundações poderiam utilizar mais efetivamente seus recursos para ajudar a transição para o acesso livre, bem como, tornar economicamente sustentável a publicação nesse novo modelo de publicação.

Já na declaração em si, publicada em 2002, é tratada a questão da “problemática da comunicação científica no contexto contemporâneo, no qual a tradição da comunicação por meio de periódicos têm seus processos influenciados pelas novas tecnologias e pela Internet” (COSTA, 2014, p. 48). Na BOAI foram propostas duas estratégias complementares para obter o acesso livre para a literatura científica: o auto-arquivamento em repositórios e a publicação em periódicos em acesso livre (SARMENTO *et al.*, 2005, p. 4), que mais tarde ficaram conhecidas como Green e Golden Road (via verde e via dourada respectivamente).

Costa (2014, p. 50) descreve a *Green Road*, Via Verde ou auto-arquivamento, como consistindo na estratégia em que “os próprios autores depositarem uma cópia de suas publicações revisadas por pares em arquivos de acesso aberto que estejam em conformidade com os padrões OAI” . Já a *Golden Road* é definida como “alteração no modelo de negócio dos periódicos já existentes e a criação de periódicos de acesso aberto, que deveriam buscar outras formas de financiamento que não fosse pela cobrança pelo acesso aos seus conteúdos” (COSTA, 2014, p. 50).

A segunda declaração, a *Bethesda Statement on Open Access Publishing* também foi baseada em uma reunião de profissionais relacionados à informação biomédica. O encontro ocorreu em Maryland, EUA, na sede do Howard Hughes Medical Institute, em 2003. Eles objetivavam discutir formas de fomentar o OA à literatura científica (SARMENTO *et al.*, 2005, p. 4). Segundo Costa e Leite essa declaração (2015, p. 3) “propôs o estabelecimento do direito ao acesso livre à informação científica por meio da atribuição de licenças de uso e distribuição” sendo então definida por essa proposta.

Vale ressaltar que nessa declaração é dada como condição para ser considerado um documento como de acesso aberto a disponibilização dos materiais suplementares, que incluiriam os dados de pesquisa.

Ao estabelecer o que seria uma publicação de acesso aberto a declaração define (COSTA, 2014, p. 51)

(...) duas condições para que uma publicação seja entendida como de acesso aberto. A primeira delas é a concessão do autor aos usuários de licença de cópia, uso, distribuição, transmissão, exibição e a produção de obras derivadas. A segunda é o arquivamento imediato de uma cópia da publicação em um repositório digital online de um instituto de pesquisa ou organização, afim de que esteja em conformidade com os padrões de interoperabilidade e preservação.

Sarmiento *et al.* (2005, p. 5) ressalta que seria necessária uma mudança na forma como se dá a publicação por pesquisadores. Enfatizando o modelo de acesso livre, o financiamento das despesas relacionadas e uma profunda alteração no método de avaliação de trabalhos por pesquisadores baseado no periódico onde foi publicado.

A terceira e final declaração, a Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, também de 2003, e foi assinada por 55 instituições. No documento é “destacada a importância da formalização de políticas de informação para o acesso aberto.” (COSTA; LEITE 2015, p. 3). Ela objetivava o fomento à publicações em acesso aberto, reforçando a transição de paradigmas, e defender seu reconhecimento em avaliações científicas.

De acordo com Sarmiento *et al.* (2005, p. 6) as duas últimas declarações, a de Bethesda e de Berlim, são complementares à primeira, a BOAI, com as duas condicionando fatores que tornam uma publicação de OA. Juntas elas formam as “três declarações que ficaram internacionalmente conhecidas como BBB” (COSTA; LEITE 2015, p. 3).

A Tabela 1 apresentada por Sarmiento *et al.* ilustra as principais características e divergências entre as três declarações aqui consideradas.

**Tabela 1-** Comparação entre os textos das declarações de Budapeste, Bethesda e Berlim que regem o Acesso Livre ao conhecimento

	<b>Declaração de Budapeste</b> Área: Genérica	<b>Declaração de Bethesda</b> Área: Biomédica	<b>Declaração de Berlim</b> Área: Ciências e Humanidades
<b>Qualidade</b> (peer review)	- artigos de periódicos revisados - pre-prints sem revisão	- artigos revisados	- aprovação pela comunidade científica (revisado)
<b>Copyright</b>	- controle sobre a integridade do trabalho - direito de propriedade intelectual - direito a citação (atribuição da autoria)	- atribuição da autoria	- atribuição da autoria
<b>Finalidade da utilização</b>	- várias, incluindo qualquer outro propósito legal	- várias, incluindo qualquer propósito responsável	- várias, incluindo qualquer propósito responsável
<b>Obras derivadas</b>	- não menciona	- permite	- permite
<b>Acesso</b>	- gratuito - público (mundial)	- gratuito - mundial - irrevogável - perpétuo (necessidade de preservação)	- gratuito - mundial - irrevogável

Fonte: Sarmiento et. al. (2005, p. 6)

A partir da tabela percebe-se que a única área especificada é a área da saúde reafirmando a sua relevância nas declarações e nas articulações iniciais do movimento de OA. Todas as três declarações também destacam a importância do acesso gratuito independente da finalidade, contanto que seja garantido o copyright devido.

Mais especificamente no Brasil, o movimento Open Access foi iniciado “com o Manifesto Brasileiro de Apoio ao Acesso Livre à Informação Científica, publicado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) em 2005 (FERREIRA, 2012, p. 32).

O próprio surgimento do movimento OA está ligado às ciências da saúde, como ressalta Giglia (2007, p. 204) ao lembrar que a declaração de Bethesda foi lançada a partir de uma reunião com pesquisadores de áreas biomédicas no Instituto Médico Howard Hughes.

O movimento OA se expande não apenas para os trabalhos e artigos em si, mas também para os dados de pesquisa coletados e utilizados na pesquisa. Fomentando o compartilhamento desses dados e aumentando a demanda de estruturas especializadas para tratar de tais dados, os repositórios de dados de pesquisa ganham espaço nestas discussões.

### 2.3. Compartilhamento de dados de pesquisa e os repositórios de dados

Há atualmente um crescente acaloramento da discussão relativa aos dados resultantes de atividades de investigação dentro do próprio sistema da comunicação científica e até mesmo no modo de funcionamento da ciência em si (RODRIGUES, et. al, 2010, p. 2).

Ferreira (2012, p. 20) define dados de pesquisa como “todas as fontes utilizadas para observação do estudo e inferência dos resultados.”. Ou ainda “todos aqueles produzidos ou coletados pelo método científico, a fim de serem utilizados para tirar conclusões para os fenômenos encontrados. “ (FERREIRA, 2012, p. 22). Ainda de acordo com a autora, eles são essenciais pois é partindo-se deles que se inferem conclusões devendo ser sempre registrados e preservados para futuras consultas.

O compartilhamento de dados então pode trazer diversos benefícios, Medeiros e Caregnato (2012, p. 316) definem que ele “possibilita que esforços sejam focados na análise dos dados já existentes e compartilhados, reduzindo custos e esforços e possibilitando avanços efetivamente mais rápidos.” e acrescentam:

A perspectiva de compartilhar dados entre cientistas que estão distribuídos por todo o mundo pode fazer com que as interações e possibilidades de colaboração se tornem cada vez mais reais entre instituições, laboratórios etc., criando ambientes via internet onde as discussões crescem e promovem o desenvolvimento da ciência e da tecnologia, permitindo que recursos sejam utilizados de forma a avançar em passos mais largos a partir do reuso de recursos financeiros, humanos, tecnológicos, entre outros, já empregados (MEDEIROS E CAREGNATO, 2012, p. 320).

Medeiros e Caregnato (2012, p. 312-312) ainda falam do potencial informático atual que vem desenvolvendo meios de compartilhamento de dados científicos primários e que cria benefícios, como o acesso a conteúdos já mapeados e corretamente anotados. Porém esse grande enfoque apenas no desenvolvimento de sistemas computacionais acaba por deixar à margem estudos teóricos e conceituais, o que tem potencial para causar confusões e conflitos dentro do processo visado. Ainda em Medeiros e Caregnato (2012, p. 315) vale lembrar que o avanço da tecnologia não instituiu o compartilhamento e exploração de dados mas sim os impulsionou.

Ferreira (2012, p. 39) comenta sobre essas mudanças e seus benefícios ao afirmar que essas possibilidades de compartilhamento e acesso a dados científicos se associam ao paradigma de acesso aberto e transparência das pesquisas, colaborando de forma a legitimar trabalhos de pesquisadores, Brody *et al.* (2007, p. 1) comentam na mesma . Afirmando que o OA pode ser estendido para os dados, mesmo que nem todos os pesquisadores queiram disponibiliza-los ou que as estruturas online de arquivamento de publicações ainda não estejam completamente funcionais para essa tarefa.

Suber (2009, p. 5) ainda ressalta que dados são fatos não passíveis de copyright e que mesmo em países que os consideram como bases de dados parcialmente passíveis de copyright, o direito pertence ao compilador, que geralmente é o pesquisador e não o editor (isso no caso de publicações como artigos em periódicos científicos, que apresentam esse papel do editor). O que faria com que raramente editores se oponham ao acesso aberto de dados de pesquisa pois isso não interfere com a venda de seu produto, pelo contrário, facilita a pesquisa e aumenta o valor final.

O movimento OA está intimamente relacionado a outro movimento conhecido então como Open Data que também aborda dados de pesquisa e de acordo com a homepage da organização Panton Principles (PANTONPRINCIPLES, tradução nossa)

Por Open Data na ciência nós queremos dizer que está livremente acessível na internet pública permitindo qualquer usuário fazer o download, copiar, analisar, reprocessar, passar em softwares ou usá-las para qualquer propósito sem barreiras financeiras, legais, ou técnicas que não aquelas inseparáveis de acessar a própria internet. Para esse fim dados relacionados à ciência publicada devem ser explicitamente postos em domínio público

Também de acordo com Rodrigues *et al.* (2010, p. 9), esse movimento dos dados científicos abertos se assemelha ao Open Source e ao Open Access pois requer que os dados estejam disponíveis gratuitamente na internet sem restrições de copyright, patente, ou outros.

O documento da OECD (ORGANIZATIONS FOR ECONOMIC CO-OPERATION DEVELOPMENT, 2007) com recomendações para o acesso aos dados de pesquisa de financiamento público, mesmo sendo voltado apenas para dados obtidos a partir de um financiamento público, apresenta importantes elementos para a discussão. Seus conceitos podem ser abstraídos para um escopo mais geral, e aqui será utilizado para comparação com os resultados encontrados na coleta.

O OECD é um fórum de governos democráticos de 30 países que trata de desafios econômicos, sociais e ambientais da globalização. O documento em questão, o OECD Principles and Guidelines for Access to Research Data from Public Funding, foi lançado em 2006 e é resultado de uma discussão em 2004 realizada em Paris. Nela foi reconhecida a importância do acesso a dados de pesquisa e a necessidade de se elaborar um guia com princípios e recomendações para facilitar e otimizar o custo benefício do acesso.

No documento foram estabelecidos 13 princípios para o acesso, foram eles: abertura, flexibilidade, transparência, conformidade legal, proteção à propriedade intelectual, responsabilidade formal, profissionalismo, interoperabilidade, qualidade, segurança, eficiência, contabilidade e por fim sustentabilidade. Segue um quadro explicativo com os princípios.

**Quadro 1-** Princípios OECD para acesso a dados de pesquisa

<b>Princípio</b>	<b>Descrição</b>
01. Abertura	Acesso igualitário para a comunidade de pesquisa internacional aos menores custos possíveis.
02. Flexibilidade	Requer levar em consideração mudanças rápidas e imprevisíveis.
03. Transparência	Disponibilização de informações sobre os dados de pesquisa e organizações que os produziram e especificações e condições de uso.
04. Conformidade legal	Acordos de acesso devem respeitar direitos e interesses legais dos envolvidos.
05. Proteção à propriedade intelectual	Acordos de acesso devem respeitar a aplicabilidade do copyright e qualquer outra lei e propriedade intelectual.
06. Responsabilidade formal	Acordos de acesso devem promover práticas institucionais explícitas e formais, ser desenvolvidos com consulta a representantes de todos envolvidos, e ser responsáveis por fatores envolvendo os dados.
07. Profissionalismo	Acordos institucionais para a manutenção de dados de pesquisa devem ser baseados em padrões profissionais relevantes e valores dos códigos de conduta das comunidades científicas envolvidas.
08. Interoperabilidade	Deve ser tanto tecnológica quanto semântica.
09. Qualidade	Assegurar a conformidade com padrões de qualidade explícitos.
10. Segurança	Uso de técnicas e instrumentos para garantir a integridade e segurança dos dados de pesquisa.
11. Eficiência	Evitar a custosa e desnecessária multiplicação de esforço na coleta de dados.
12. Contabilidade	A performance dos acordos de acesso deve ser avaliada periodicamente.
13. Sustentabilidade	Tomar responsabilidades administrativas sobre as medidas para garantir o acesso permanente a dados que necessitam de retenção a longo prazo.

Fonte: Organizations for economic co-operation development (2007, adaptado pelo autor)

A expectativa é que a aplicação dos 13 princípios para acesso a dados de pesquisa, possibilitaria uma base consistente para desenvolver um repositório de dados de pesquisa ou para adaptar um repositório existente.

Medeiros e Caregnato (2012, p. 319) apontam que “a partir da preservação e manutenção dos registros existentes em um centro, seja possível agregar valor aos dados de pesquisa e permitir seu compartilhamento e reuso” destacando a importância da curadoria. Já Rodrigues *et. al.* (2010, p.17) afirmam que é crescente o número de repositórios institucionais e que eles poderiam atuar na curadoria de dados. Porém suas funcionalidades com relação à especificidades dos dados de pesquisa ainda são reduzidas, ou seja a atuação desses repositórios quanto a dados de pesquisa ainda deixa a desejar.

Ainda em Rodrigues *et al.* (2010, p. 15) a necessidade de preservação de dados é comum a todas as áreas disciplinares, porém o tipo de dado e o período de preservação são variados. Ou seja as estratégias de preservação são dependentes das características dos dados e da área do conhecimento.

Os autores ainda apontam que para que o compartilhamento e reuso possam ocorrer é essencial que organizações e indivíduos sejam responsáveis pela sua curadoria e não somente os próprios pesquisadores (RODRIGUES *et al.*, 2010, p. 17). Vale ressaltar então que no repositório de dados “pode ir-se muito além desta visão de repositório de objectos, uma vez que cada conjunto de dados tem características próprias e por isso pode requerer um tratamento diferenciado.” (RODRIGUES *et. al.*, 2010, p. 23).

Todos esses fatores apontam para a necessidade de repositórios de dados, porém mesmo com a problemática da curadoria de dados ser reconhecida e diversas iniciativas desenvolvidas, ela geralmente ainda não recebe a atenção necessária.

Para a análise do desenvolvimento de repositórios de dados de pesquisa Rodrigues *et al.* (2010, p. 23) definem 3 características essenciais para um repositório de dados.

- 1- A garantia de um armazenamento fiável com manutenção e garantias, que quer dizer que um repositório de dados deve ter ferramentas e políticas confiáveis e sólidas, evitando a perda de dados.
- 2- Uma descrição do conjunto de dados e metadados sobre o mesmo, que quer dizer um eficiente conjunto de campos de descrição a fim de representar e auxiliar a busca do dado, e por fim
- 3- A representação do item que depende da natureza do conjunto, que seria a forma como o dado é apresentado

Eles ainda afirmam que seria tarefa dos responsáveis pelo repositório prover apoio à curadoria e partilha e disponibilizar informações sobre o acesso e utilização dos conjuntos de dados. Com isso os autores alertam que há muito a fazer tanto no estabelecimento de práticas de descrição adequada como na normalização dos descritores a usar” (RODRIGUES *et al.*, 2010, p. 28).

## **2.4 Ciências da saúde no open access**

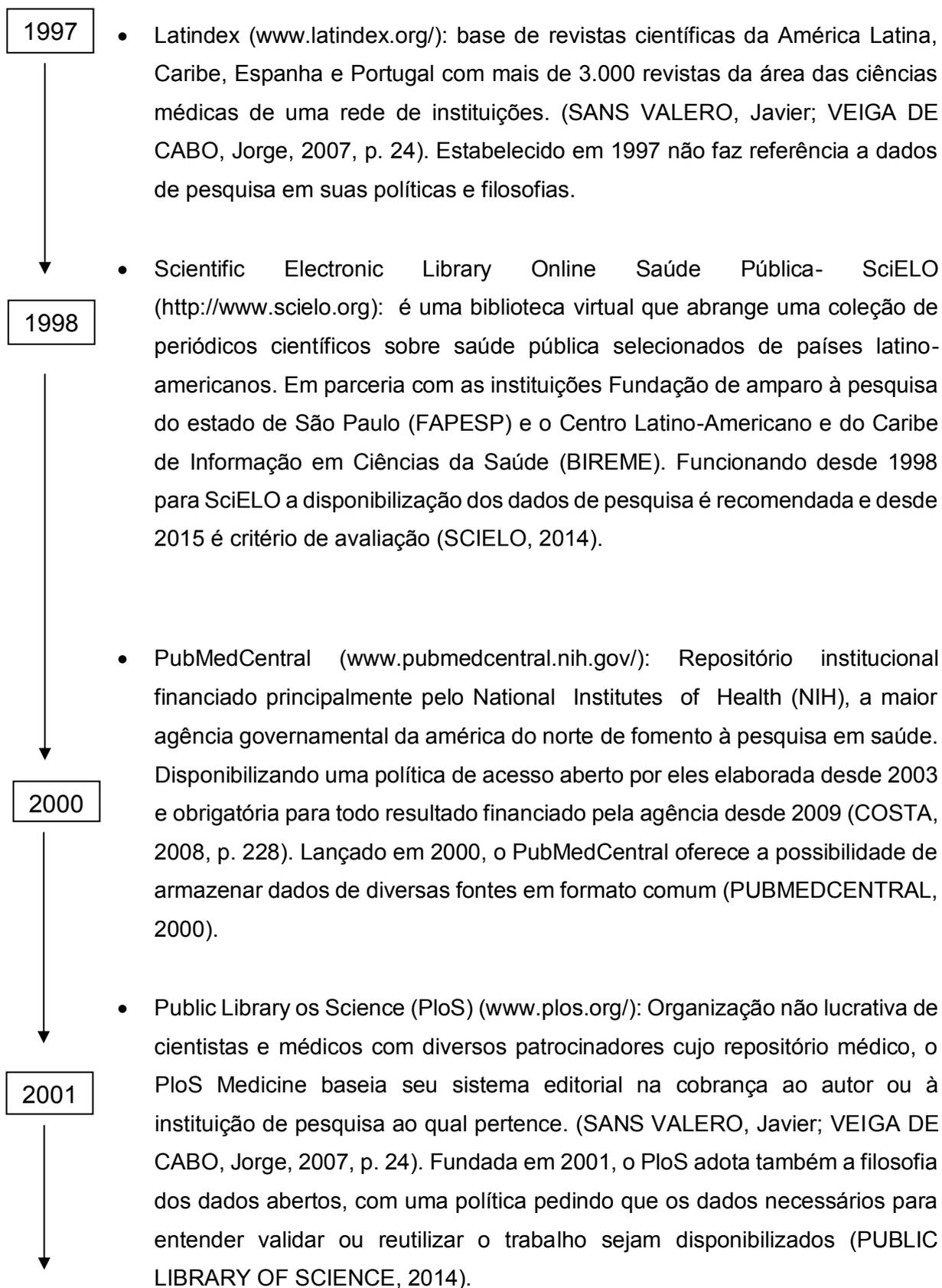
Como já citado anteriormente o movimento OA teve seu surgimento intimamente relacionado às ciências da saúde. Pode-se abstrair então que a percepção nas declarações das necessidades recém surgidas, como o acesso sem custos, irrestrito e online, advém de

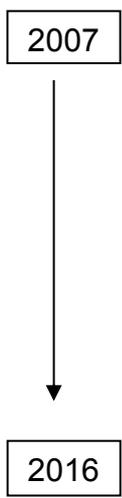
uma percepção das próprias necessidades das ciências da saúde de uma forma geral, que poderiam ser estendidas para outras áreas. Pode-se perceber então que a preocupação com a disseminação de dados de pesquisa já está presente nas ciências da saúde desde antes da elaboração das declarações.

Nas áreas das ciências da saúde encontramos uma particularidade, pois existem diversas limitações quanto a salvaguarda de dados e sua disponibilização. Isso porque muitas vezes pode gerar constrangimentos para pacientes (RODRIGUES *et al.* 2010, p. 16). Glover, *et al.* (2006, p. 197, tradução nossa) também ressalta dificuldades, porém com relação à publicação em OA

O modelo vigente de longa data que serviu a sociedade acadêmica na comunicação científica está sobre crescente pressão a medida que as assinaturas dos periódicos continuam a subir de preço acima da inflação. No campo das ciências bio-médicas vários princípios apoiam a ordem existente, particularmente, o processo de validação pelos pares e o prestígio acadêmico associado ao local de publicação.

Porém o autor lembra que no movimento OA, mesmo com suas diversas opiniões no mercado editorial da área das ciências da saúde, ainda prevalece a ideia geral que um artigo deve ser livremente acessível ao leitor. Sendo que o método mais empregado é aquele onde o autor ou a instituição de pesquisa o financiam (GLOVER *et. al.*, 2006, p. 199). Harnad (2007, p. 3) também comenta sobre o assunto ao definir o “*Acesso dos pagantes de impostos*”, onde ao combater o lobby dos editores contra a via verde do OA, foi formada uma aliança, focada basicamente em biomedicina e seus potenciais benefícios à saúde. Outras diversas iniciativas de acesso aberto abrangendo áreas da saúde também se destacam. Seguem algumas delas organizadas historicamente:



- 
- Health InterNetwork Access to Research Initiative (HINARI) ([www.who.int/hinari/es/](http://www.who.int/hinari/es/)): coleção de literatura biomédica e de saúde voltada para países em desenvolvimento, com mais de 3.070 revistas disponíveis para instituições de saúde em 113 países. (SANS VALERO, Javier; VEIGA DE CABO, Jorge, 2007, p. 24). Lançado em 2002, também não faz referência a dados de pesquisa.
  - Zika Open ou Acordo Zika ([http://www.who.int/bulletin/online\\_first/zika\\_open/en/](http://www.who.int/bulletin/online_first/zika_open/en/)): O Boletim da organização mundial da saúde decidiu implementar novos protocolos de partilha e relato de dados, estabelecidos especificamente para tratar da escassez de informações e dados existente sobre a atual epidemia do Zika vírus. Ele disponibiliza dados para uso, distribuição e reprodução em qualquer meio gratuitamente sendo aderido por cientistas e 30 publicações científicas internacionais, sendo a única Brasileira participante a Fiocruz (Fundação Oswaldo Cruz).

Matsubayashi et. al. (2009, p. 6) fizeram um estudo publicado em 2009 com medições sobre o status de acesso livre no campo biomédico sobre o ano de 2005. Nele foram coletados em 2006, 4667 artigos de forma aleatória publicados entre janeiro e setembro de 2005, com entre 11 e 19 páginas, do repositório PubMed escolhido por sua ampla abrangência e popularidade no campo biomédico. Como resultado obtiveram a tabela e o gráfico a seguir.

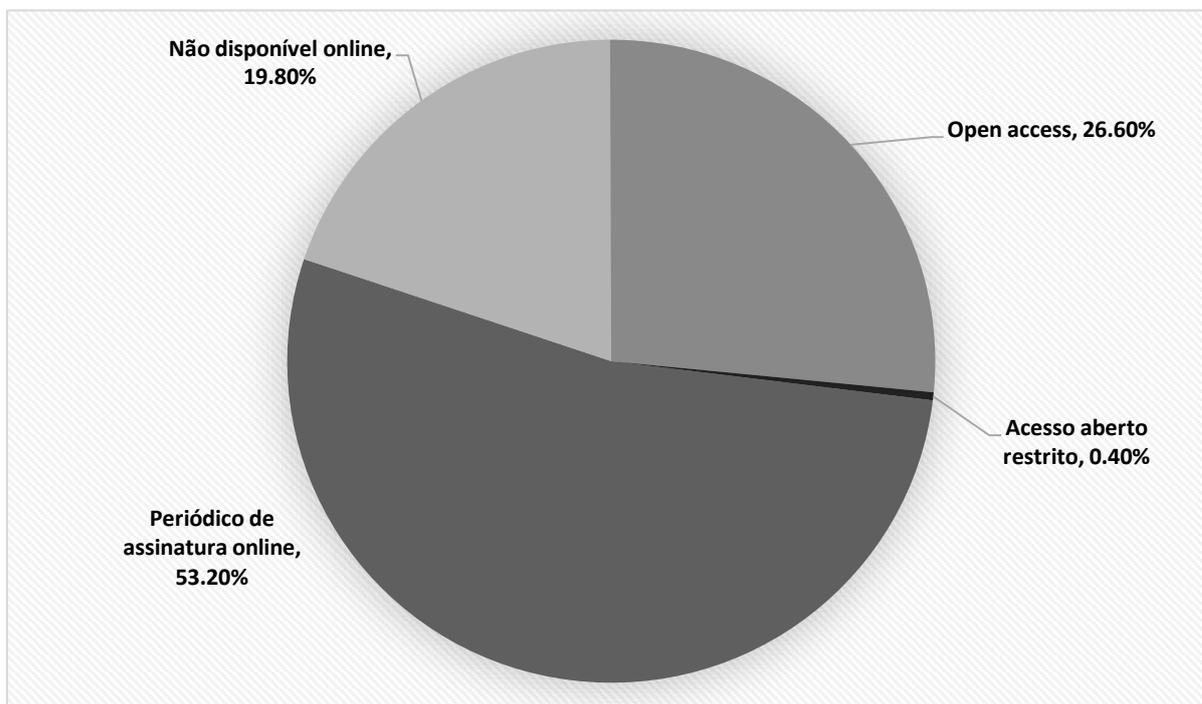
**Tabela 2-** Distribuição de artigos de acesso aberto (OA) por país do primeiro autor\*

País	Nº	% de arts. publicados em OA	Distribuição de artigos AO				
			PubMed Central	Sites de periódicos	Repositórios institucionais	Sites dos autores	Plataforma ou portal de periódico
1. Estados Unidos	1,241	30.1%	32.4%	72.2%	7.8%	5.4%	15.3%
2. Reino Unido	312	24.0%	37.3%	77.3%	8.0%	10.7%	4.0%
3. Japão	234	20.9%	10.2%	57.1%	6.1%	2.0%	40.8%
4. Alemanha	212	18.9%	37.5%	75.0%	5.0%	12.5%	-
5. China	154	16.9%	19.2%	92.3%	-	-	15.4%
6. Canada	149	37.6%	37.5%	78.6%	1.8%	8.9%	8.9%
7. Itália	133	20.3%	33.3%	81.5%	11.1%	-	3.7%
8. França	130	22.3%	41.4%	89.7%	6.9%	3.4%	-
9. Austrália	101	32.7%	42.4%	75.8%	12.1%	9.1%	3.0%
10. Holanda	87	26.4%	34.8%	87.0%	8.7%	4.3%	-
11. Espanha	81	24.7%	10.0%	85.0%	5.0%	-	5.0%
12. Suécia	67	29.9%	55.0%	75.0%	15.0%	5.0%	-
13. Suíça	61	18.0%	27.3%	72.7%	-	18.2%	-
14. Índia	60	40.0%	33.3%	87.5%	4.2%	-	12.5%
15. Brasil	55	36.4%	5.0%	10.0%	5.0%	-	85.0%
16. Turquia	51	25.5%	7.7%	100.0%	-	-	15.4%
17. Polônia	42	23.8%	10.0%	80.0%	10.0%	10.0%	-
18. Bélgica	36	41.7%	20.0%	80.0%	-	-	6.7%
19. República da Coreia	30	30.0%	11.1%	77.8%	-	-	11.1%
20. Taiwan	30	20.0%	16.7%	83.3%	-	16.7%	-

\*Top vinte países em termos de número total de artigos publicados

Fonte: Matsubayashi et. al. (2009, p. 7, tradução nossa)

A partir da Tabela 2, percebe-se uma predominância do hemisfério norte com relação à quantidade de artigos e quanto à sua presença no PUBMED, porém ao se tratar da porcentagem de artigos em OA essa predominância desaparece e outros países como a Índia e o Brasil também se destacam juntamente com a Bélgica, o Canadá e outros.

**Gráfico 1-** Disponibilidade do texto dos artigos selecionados (n=4,667)

Fonte: Matsubayashi et. al. (2009, p. 8, tradução nossa)

A partir dos dados apresentados percebemos que a grande maioria dos artigos são publicados nos Estados Unidos e que a porcentagem de artigos publicado em OA segue um padrão por volta dos 30%. Percebe-se também que uma porcentagem similar encontra-se no PubMedCenter, com algumas exceções, como o Japão, Espanha, e outros. Uma grande porcentagem, por volta dos 80%, dos artigos encontram-se em sites de periódicos, uma porcentagem baixa ou inexpressiva se encontra em repositórios institucionais e em sites de autores e em plataformas ou portais de periódicos, este último com exceção do Japão e do Brasil.

Além das iniciativas de acesso aberto dentro da área das ciências da saúde vale acrescentar a BioMed Central ([www.biomedcentral.com/](http://www.biomedcentral.com/)), editora do Reino Unido, voltada para o acesso livre e imediato à pesquisa biomédica revisada por pares contém mais de 140 revistas sobre ciências da saúde. (SANS VALERO; VEIGA DE CABO, 2007, p. 23). Essa iniciativa se destaca especialmente por sua política editorial onde “A apresentação de um manuscrito para um periódico BioMed Central implica que materiais descritos no manuscrito, incluindo todo dado bruto relevante, estará disponível gratuitamente para qualquer cientista que deseje usar-los.” (BIOMED CENTRAL, tradução nossa). Ainda de acordo com sua política a BioMed Central acredita que conjuntos de dados em que são baseadas as conclusões devem ser depositadas pelo autor em repositórios de dados disponíveis ao público ou apresentados no próprio documento principal.

Pode-se perceber então, a partir do surgimento e adaptação dessas e outras diversas iniciativas na área das ciências da saúde, uma crescente preocupação voltada para suprir a demanda não somente de acesso aberto mas também com relação à disponibilização dos dados de pesquisa.

Com as rápidas mudanças de paradigma envolvendo a comunicação científica, como a evolução tecnológica e novos modelos organizacionais da mesma, foi surgindo a proposta do movimento OA. Que se propõe a disponibilizar o acesso livre de custos e outras restrições ou barreiras como impedições por direitos autorais ou de posse especialmente no meio digital. Essa proposta OA não só veio dessas mudanças na comunicação científica como também foi uma resposta às novas demandas que surgiam.

Uma dessas novas demandas é a disponibilização e compartilhamento não somente do documento do trabalho em si, mas também os dados de pesquisa coletados necessários para o entendimento, expansão e recriação do trabalho.

As ciências da saúde sempre foram uma área proeminente ao se tratar de avanços em diversos sentidos, e com relação ao movimento open access não foi diferente. Participando ativamente no desenvolvimento da declaração de Bethesda e servindo como vetor de impulso para o movimento como um todo, percebe-se nela já uma crescente preocupação com a disseminação dos dados de pesquisa como visto nas iniciativas previamente citadas. Logo uma futura tendência é surgimento de repositórios especializados em dados de pesquisa ou a adaptação dos repositórios atuais para abranger esses dados. Na próxima seção encontraremos a metodologia utilizada neste trabalho.

### **3. Metodologia**

O propósito desse estudo foi descritivo, 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. ” Correspondendo aos objetivos já estabelecidos de levantar, analisar e avaliar o conjunto de repositórios de dados de pesquisa na área das ciências da saúde.

#### **3.1. Universo e amostra**

O universo em questão são os repositórios de dados de pesquisa nas ciências da saúde. Para a definição de uma amostra foram selecionados todos os 20 repositórios de dados online e certificados encontrados a partir da pesquisa, que estavam cadastrados no Registry of Research Data Repositories- RE3DATA (<http://www.re3data.org/>). A amostra foi selecionada a partir da seguinte estratégia de busca.

- Selecionada a opção “*Search*” (busca)
- Entre as opções de filtro foi selecionada a opção “*Subjects*” (ou disciplinas temáticas)
- Foi selecionado o tópico “*Basic Biological and Medical Research*” (Ciências médicas e da saúde) para seleção de repositórios que possuem inscritos tal disciplina
- Dos 303 resultados foram refinados apenas os 20 que possuíam algum tipo de certificação, como por exemplo o Trusted Digital Repository ou o CLARIN certificate B que estabelecem padrões e selos de qualidade

A busca foi realizada em março de 2015, como parte do projeto do Programa de Atividade Complementar (PAC) intitulado “Mapeamento e análise de repositórios de dados de pesquisa”.

#### **3.2. Fontes de informação e técnica de coleta de dados**

Como principal fonte de informação foi escolhido o site de registro global de repositórios de dados de pesquisa RE3DATA. O diretório é um registro internacional de repositórios de dados de pesquisa que abrange repositórios de diferentes áreas acadêmicas e os disponibiliza para o arquivamento permanente para o acesso a conjuntos de dados.

O procedimento de coleta de dados foi a partir do próprio site RE3DATA na opção *Search* (busca), seleção do filtro *Subjecs* (assuntos), e escolha da tag *Basic Biological and Medical Research* (ciências médicas e da saúde). Foram selecionados os 20 repositórios que são certificados.

### 3.3. Técnica de análise dos resultados

A técnica de análise dos resultados foi descritivo. Em seguida os repositórios foram analisados em 3 grandes categorias desenvolvidas a partir da adaptação das 3 características essenciais para um repositório de dados proposto por Rodrigues *et al.* (2010, p. 23). Que são:

- 1- A garantia de um armazenamento fiável com manutenção e garantias,
- 2- Uma descrição do conjunto de dados e metadados sobre o mesmo e,
- 3- A representação do item que depende da natureza do conjunto.

O primeiro aspecto abordado é a identificação do repositório com finalidade de possibilitar separações e categorizações.

**Tabela 3-** Identificação

Identificação	URL: link de acesso ao repositório
	Tipo de repositório: se é um repositório temático, institucional, regional, ou outro
	Instituição: o nome da instituição responsável pelo repositório
	País: localidade

Fonte: Elaboração própria.

As tabelas com as dimensões da análise que serão apresentadas foram desenvolvidas baseadas nas 3 categorias propostas por Rodrigues *et al.*, com o intuito de melhor definir e visualizar as dimensões. Foi desenvolvido também um quadro detalhado com os resultados dessa análise para cada repositório de dados estudado que se encontra nos apêndices.

Já a primeira grande categoria analisada foi a garantia de um armazenamento fiável com manutenção e garantias, ou seja armazenamento. Vale destacar que as subcategorias (software, sistema de preservação, etc) tanto dessa categoria quanto das próximas foram desenvolvidas a partir da disponibilização de informações dos próprios repositórios e a partir da experiência com o PAC com algumas adaptações e elaborações próprias.

**Tabela 4-** Armazenamento

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

Fonte: Elaboração própria.

A próxima categoria, baseada na descrição do conjunto de dados e metadados sobre o mesmo, é a descrição.

**Tabela 5-** Descrição

Descrição	Metadados (padrão): qual padrão de metadados utilizado.
	Campos de descrição: para análise desse aspecto foram definidas 3 características chave: identificação (autoria, instituição, etc.), descrição dos dados e permissões, sendo analisadas sua presença e extensão e destacadas anomalias.

Fonte: Elaboração própria.

A terceira e última grande categoria, a baseada na representação do item, ou como que ele é visualmente apresentado, e que depende da natureza do conjunto, foi denominada apresentação dos itens.

**Tabela 6-** Apresentação dos itens

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.

Como parâmetro de comparação também foi utilizado o documento Principles and Guidelines for Access to Research Data from Public Funding do OECD e seus 13 princípios já abordados.

A seguir é apresentado um quadro explicativo resumindo e ilustrando a metodologia utilizada a partir dos objetivos, universo, amostra, fonte, técnica de coleta e método de análise utilizados.

**Quadro 2-** Quadro da metodologia

<b>Objetivos específicos</b>	<b>Universo</b>	<b>Amostra</b>	<b>Fonte</b>	<b>Técnica de coleta</b>	<b>Método de análise</b>
<b>Levantar os repositórios existentes na área de ciências da saúde.</b>	Repositórios de dados de pesquisa nas ciências da saúde	Repositórios de dados de pesquisa nas ciências da saúde cadastrados no RE3DATA	RE3DATA	Pesquisa documental a partir dos repositórios	Análise descritiva
<b>Avaliar os repositórios de acordo com metodologia já existente.</b>	Repositórios de dados de pesquisa nas ciências da saúde cadastrados no RE3DATA	Repositórios de dados de pesquisa nas ciências da saúde cadastrados no RE3DATA	RE3DATA	Pesquisa documental a partir dos repositórios	Análise descritiva
<b>Análise comparativa dos repositórios.</b>	Repositórios de dados de pesquisa nas ciências da saúde cadastrados no RE3DATA	Repositórios de dados de pesquisa nas ciências da saúde cadastrados no RE3DATA	Resultados anteriores	Pesquisa documental a partir dos repositórios	Análise comparativa

Fonte: Elaboração própria

A próxima seção, Discussão dos resultados, traz as características dos repositórios de dados de pesquisa utilizados na pesquisa, bem como uma análise desses dados.

## 4. Discussão dos resultados

Nesse tópico serão apresentados os resultados obtidos a partir da metodologia aplicada. Inicialmente serão apresentados os repositórios de dados de pesquisa que foram identificados de acordo com os critérios propostos, seguido de uma descrição da amostragem encontrada, assim como uma avaliação e análise da mesma. O conjunto de repositórios de dados de pesquisa foi descrito sistematicamente de forma a permitir uma análise comparativa entre os sistemas.

### 4.1. Descrição e análise

Nesta seção a amostragem será descrita de acordo com os critérios estabelecidos e depois analisados comparativamente entre si utilizando os conceitos encontrados nos documentos já citados na metodologia.

#### 4.1.1. Identificação

Os repositórios na área de ciências da saúde certificados foram levantados no primeiro semestre de 2015, no RE3DATA. Primeiramente foi selecionado o termo ciências médicas e da saúde no campo disciplinas temáticas e logo após foi refinada a busca apenas a repositórios certificados. A amostragem dos repositórios de dados de pesquisa e seus respectivos *links* bem como suas instituições mantenedoras são:

1.HomoMINT- <http://mint.bio.uniroma2.it/HomoMINT/Welcome.do> (Universita degli Studi di Roma 'Tor Vergata')

2.American Type Culture Collection- <http://www.lgcstandards-atcc.org/> (LGC Standards)

3.MaizeGDB- <http://www.maizegdb.org/> (Iowa State University, Department of Agronomy)

4.ArrayExpress- <http://www.ebi.ac.uk/arrayexpress/> (EMBL-EBI, Functional Genomics Group)

5.Bacterial Carbohydrate Structure DataBase- <http://csdb.glycoscience.ru/bacterial/index.html> (The President of the Russian Federation Grants Council)

6. InnateDB- <http://www.innatedb.com/> (Simon Fraser University, Fiona Brinkman Laboratory)

7. mentha- <http://mentha.uniroma2.it/> (University of Tor Vergata, Department of Biology, Molecular Genetics Group)

8. INDEPTH Data Repository- <http://www.indepth-ishare.org/index.php/home> (INDEPTH Network)

9. Human Proteinpedia- <http://www.humanproteinpedia.org/> (Institute of Bioinformatics)

10. Hardin.MD- <http://hardinmd.lib.uiowa.edu/> (The University of Iowa Libraries)

11. Bii- <https://github.com/ISA-tools/BioInvIndex> (Leibniz Institute of Plant Biochemistry, Research Group Mass spectroscopy and Bioinformatics)

12. Canadian Epigenetics, Environment and Health Research Consortium Platform- <http://www.epigenomes.ca/> (Canadian Institutes of Health Research, The Canadian Epigenetics, Environment and Health)

13. Centers for Disease Control and Prevention, Data & Statistics- <http://www.cdc.gov/DataStatistics/> (Centers for Disease Control and Prevention)

14. ClinicalTrials.gov- <http://www.clinicaltrials.gov/> (HHS.gov)

15. Collaborative Psychiatric Epidemiology Surveys- <http://www.icpsr.umich.edu/icpsrweb/CPES/> (National Institute of Mental Health)

16. Danish Data Archive- <https://www.sa.dk/en/services/danish-data-archive> (Danish Data Archives)

17. DiversityData.org- <http://www.diversitydata.org/> (Brandeis University, The Heller School for Social Policy and Management, Institute for Child, Youth and Family Policy)

18. Domino- <http://mint.bio.uniroma2.it/domino/search/searchWelcome.do> (Universita degli Studi di Roma 'Tor Vergata', Dipartimento di Biologia)

19.NeuroMorpho- <http://neuromorpho.org/> (George Mason University, Krasnow Institute for Advanced Study, The Center for Neural Informatics, Neural Structures, and Neural Plasticity)

20.Neuroscience Information Framework- <http://neuinfo.org/> (Neuroscience Information Framework)

Os repositórios acessados se encontram nos mais diversos estados. Apresentam pouca ou nenhuma interoperabilidade entre si, com exceção dos que se relacionam ao PubMedCentral (Bacterial Carbohydrate Structure DataBase, Human Proteinpedia, Collaborative Psychiatric Epidemiology Surveys, NeuroMorpho, Neuroscience Information Framework), que serve como um padronizador em alguns aspectos. De forma geral os repositórios oferecem desde dados singulares, a grandes conjuntos de dados. A facilidade de navegação também foi diversificada, predominantemente nos repositórios que apresentam dados mais técnicos e específicos a navegação foi pouco intuitiva e problemática. Onde dois repositórios se destacam por problemas nestas condições: o Domino e o Danish Data Archive. A grande maioria dos repositórios oferta os dados de forma gratuita, e um em particular oferece até mesmo a compra de materiais genéticos associados aos dados, o American Type Culture Collection.

Com relação ao documento OECD Principles and Guidelines for Access to Research Data from Public Funding que estabelece os 13 princípios foi utilizado como parâmetro de comparação com os repositórios encontrados. Seu emprego teve por finalidade a definição do estado geral que se encontram esses repositórios.

Podemos perceber a partir do documento que o segundo princípio, a flexibilidade também pode ser considerado como alcançado, devido as diversas iniciativas que rapidamente surgem e se adaptam para suprir as demandas dos dados de pesquisa. E que Para o oitavo princípio, interoperabilidade, há uma situação ambivalente, com parte dos repositórios apresentando uma interoperabilidade a partir do PUBMED e outra parte não. O nono princípio, qualidade, se encontra em ótimo estado, com a grande maioria dos repositórios sendo integrada aos padrões do PUBMED ou com padrões de qualidade próprios estabelecidos.

Todos os repositórios de dados de pesquisa selecionados, apresentavam uma página com interface com ao menos a possibilidade de leitura em Inglês. Apenas um era originalmente em dinamarquês. Deles, apenas HomoMINT, Mentha e Domino, pertencem à uma mesma instituição. As instituições responsáveis pelos repositórios foram predominantemente universidades e grupos de pesquisa governamentais e particulares. Dos

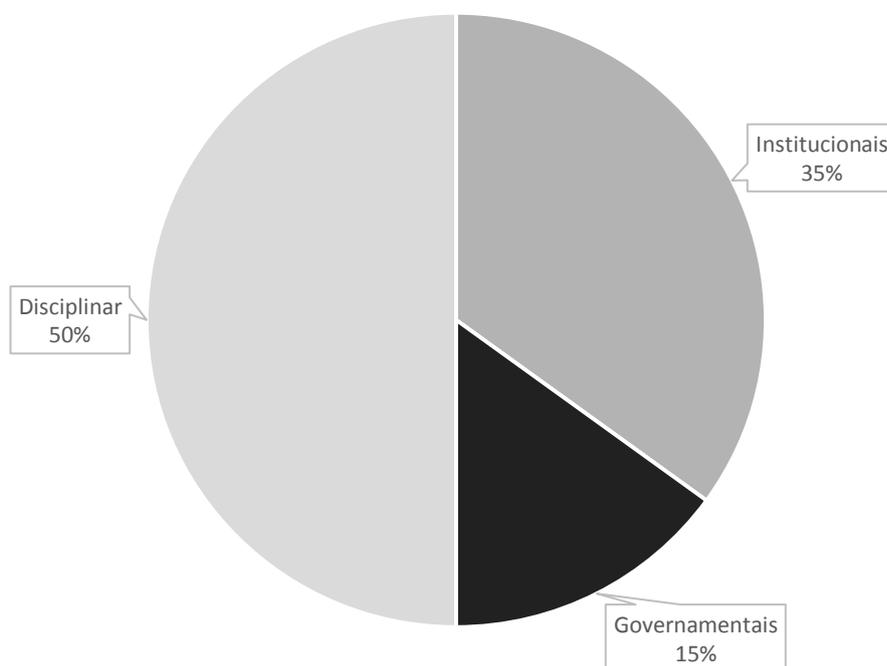
quais 8 são mantidos por 5 universidades diferentes, 5 são mantidos por organizações governamentais e 7 instituições privadas.

No quesito identificação os repositórios foram categorizados por tipo como:

- Institucionais: MaizeGDB, InnateDB, Mentha, Hardin.MD, Bii, Danish Data Archive e NeuroMorpho,
- Governamentais: ArrayExpress ,Bacterial Carbohydrate Structure DataBase e Neuroscience Information Framework e,
- Disciplinares: Centers for Disease Control and Prevention, Data & Statistics, HomoMINT, American Type Culture Collection, INDEPTH Data Repository, Human Proteinpedia, Canadian Epigenetics, Environment and Health Research Consortium Platform, ClinicalTrials.gov, Collaborative Psychiatric Epidemiology Surveys, DiversityData.org e Domino.

O tipo “Governamentais” foi adicionado posteriormente para destacar uma diferença entre repositórios governamentais e institucionais, principalmente com relação ao público alvo. O tipo de repositório mais expressivo no conjunto são os repositórios disciplinares representando 50% da amostra, ou seja metade dos 20 encontrados (Gráfico 2). Vale destacar que são considerados governamentais os repositórios de dados mantidos por algum órgão especializado ou não proveniente diretamente de um governo.

Os repositórios disciplinares encontrados foram em sua maioria voltados a dados específicos, como por exemplo material genético. Já os governamentais e institucionais, apresentaram conjuntos de dados mais diferenciados, como por exemplo estatísticas ou até mesmo imagens.

**Gráfico 2- Tipo**

Fonte: Elaboração própria

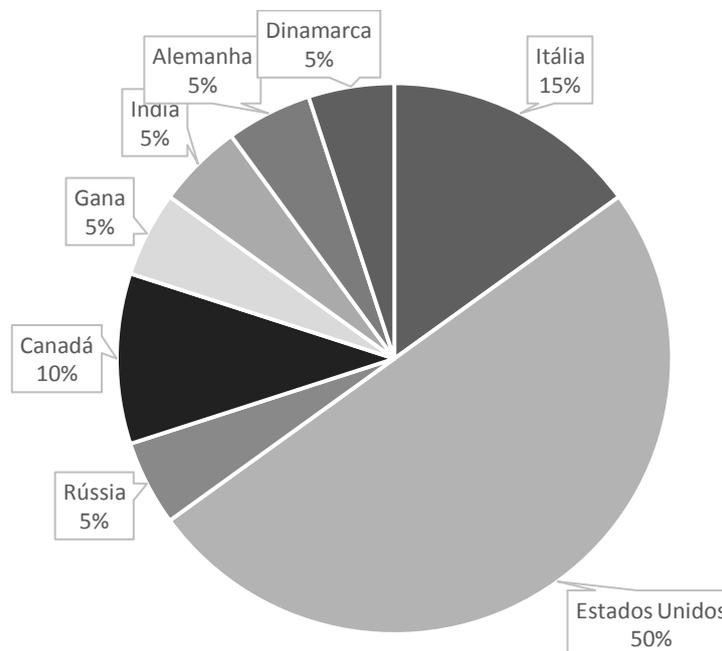
Como pode-se perceber a partir do gráfico a maior parte dos repositórios de dados de pesquisa são disciplinares com metade dos repositórios (50%), seguido de institucional com 7 repositórios (35%) e depois governamentais com 3 (15%, a menor porcentagem da amostragem no quesito).

Ainda na identificação os países de origem dos repositórios encontrados foram:

- Itália: HomoMINT, Mentha e Domino,
- Estados Unidos da América: American Type Culture Collection, MaizeGDB, Hardin.MD, Centers for Disease Control and Prevention, Data & Statistics, ClinicalTrials.gov, Collaborative Psychiatric Epidemiology Surveys, DiversityData.org, NeuroMorpho e Neuroscience Information Framework,
- Reino Unido: ArrayExpress,
- Rússia: Bacterial Carbohydrate Structure DataBase,
- Canadá: InnateDB e Canadian Epigenetics, Environment and Health Research Consortium Platform,
- Gana: INDEPTH Data Repository,
- Índia: Human Proteinpedia,
- Alemanha: Bii e,

- Dinamarca: Collaborative Psychiatric Epidemiology Surveys.

**Gráfico 3- Países**



Fonte: Elaboração própria

A partir do Gráfico 3 podemos perceber que metade dos repositórios é proveniente dos Estados Unidos, com 3 repositórios Italianos, 2 Canadenses e 1 de cada outra nacionalidade encontrada. Percebe-se então uma predominância dos países do Norte (América do Norte e Europa). Apenas 10% são de outras regiões e nenhum da América Latina.

#### 4.1.2. Armazenamento

Em relação a primeira categoria “armazenamento”, os únicos repositórios que disponibilizavam informações sobre os *software* utilizados foram:

- o HomoMINT com o MINT Viewer,
- o Bacterial Carbohydrate Structure DataBase com o SOAP,
- o InnateDB com o Cerebral program (Java), o ProbeLynx e o INVEX,
- o Mentha com o REST,
- o Bii com o BioInvIndex
- e por fim o Canadian Epigenetics, Environment and Health Research Consortium Platform com OpenCEM como ilustrado no Quadro 3.

Quadro 3- Software

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

Fonte: Elaboração própria

O HomoMINT com seu software MINT viewer<sup>2</sup> permite uma visualização gráfica no contexto de dados de alto rendimento. Já o software<sup>3</sup> 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. Sobre os softwares<sup>4</sup> do InnateDB, o Cerebra program é um plugin Java para visualização de dados sobre moléculas, o INVEX para análise e visualização de parâmetros e metadados desses dados e o ProbeLynx para anotações sobre microarranjos. O software REST<sup>5</sup> do Mentha é uma interface que oferece uma aplicação gráfica e ferramentas que representam interações e caminhos. O BiolnvIndex<sup>6</sup> do Bii é uma aplicação da web, modelo de base de dados persistente e pacote de serviços e web serviços. Por fim o OpenCent<sup>7</sup> é um software de análise, manuseio e publicação de dados epigenéticos.

#### 4.1.3. Preservação

Poucos repositórios apresentam informações referentes aos seus sistemas de preservação. Os repositórios que abordavam o tópico eram:

- o MaizeGDB que se utiliza de curadoria comunitária, que é a incorporação de dados já curados por outras iniciativas,
- o InnateDB e o Mentha que usam curadoria manual, que é aquela feita pelo próprio pesquisador pessoalmente, e
- o NeuroMorpho com curadoria centralizada que é aquela feita exclusivamente por uma equipe dentro da própria instituição.

<sup>2</sup> Fonte: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1751541/>

<sup>3</sup> Fonte: <http://csdb.glycoscience.ru/bacterial/index.html>

<sup>4</sup> Fonte: <http://www.innatedb.com/redirect.do?go=resourcesSw>

<sup>5</sup> Fonte: <http://mentha.uniroma2.it/developers.php>

<sup>6</sup> Fonte: <https://github.com/ISA-tools/BiolnvIndex>

<sup>7</sup> Fonte: <http://www.epigenomes.ca/tools-and-software>

Vale destacar que a curadoria em si abrange outras práticas que vão além da preservação e conservação. Além disso o único repositório que informa utilizar tecnologia URL persistente é o Bii. O Quadro 4 abaixo ilustra a situação encontrada.

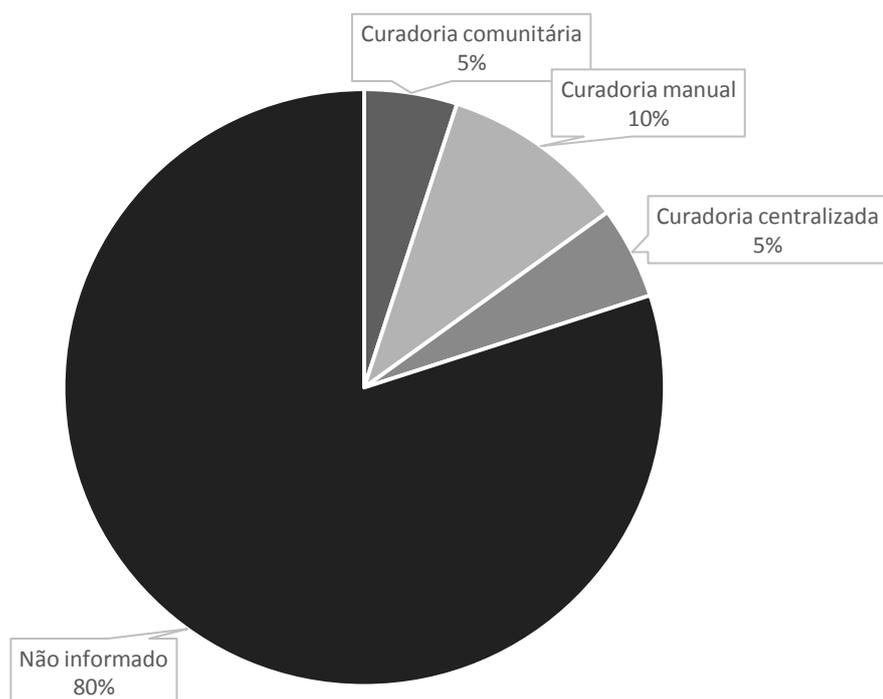
**Quadro 4-** Sistemas de preservação

<b>Repositório</b>	<b>Sistema de preservação</b>
MaizeGDB	Curadoria comunitária
InnateDB	Curadoria manual
Mentha	
Neuromorpho	Curadoria centralizada
Bii	URL persistente

Fonte: Elaboração própria

Vale ainda destacar que boa parte dos repositórios de dados não apresentou informações sobre seu sistema de preservação (Gráfico 4), indicando algumas possibilidades como, não possuir, não haver preocupação com a transparência ou não apresentar comprometimento com a própria prática da preservação.

Logo, em valores numéricos, a grande maioria dos repositórios (16) não informa se utiliza algum tipo de método de preservação, apenas 2 repositórios dizem usar curadoria manual, 1 curadoria comunitária e apenas 1 outro curadoria centralizada. O que demonstra que ainda não há atualmente a preocupação generalizada devida com a preservação dos dados de pesquisa.

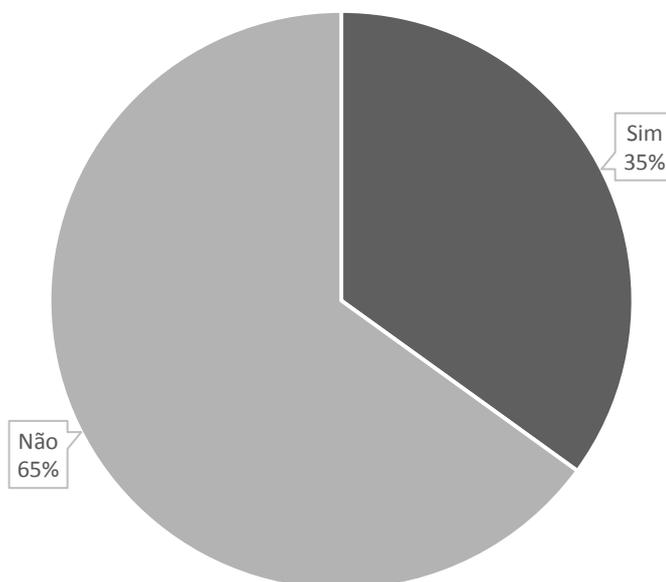
**Gráfico 4- Sistema de preservação**

Fonte: Elaboração própria

#### 4.1.4. Referência a instituição do pesquisador

Os repositórios que fazem algum tipo de referência a instituição do pesquisador são:

- Bacterial Carbohydrate Structure DataBase,
- INDEPTH Data Repository,
- Hardin.MD,
- ClinicalTrials.gov,
- Collaborative Psychiatric Epidemiology Surveys e,
- Danish Data Archive.

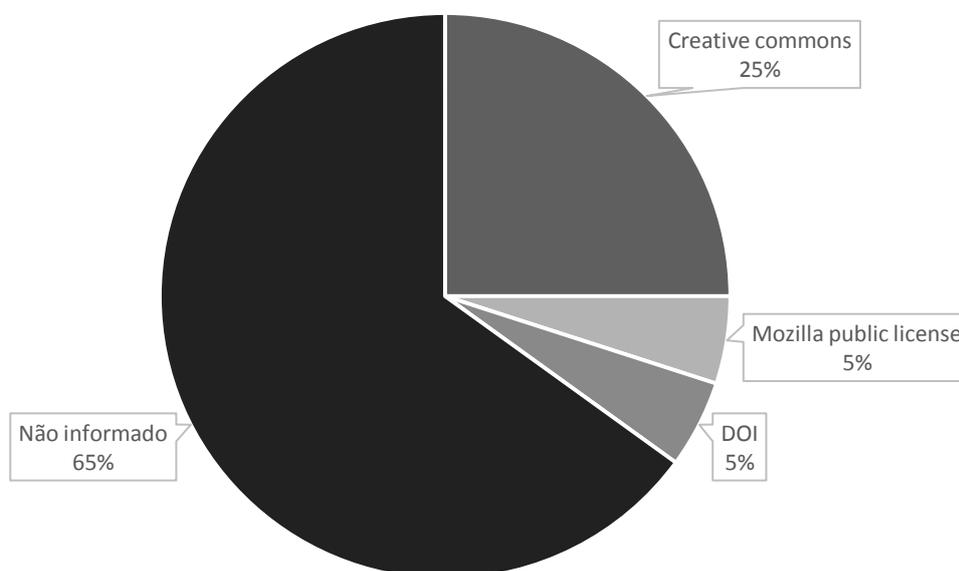
**Gráfico 5- Instituição do pesquisador**

Fonte: Elaboração própria

Com o auxílio do Gráfico 5 podemos perceber que a maioria dos repositórios não informam sobre a instituição do pesquisador (13 repositórios), enquanto apenas 7 repositórios apresentam algum tipo de informação ou indicação sobre a mesma. Isso pode apontar para uma falta de preocupação da instituição de origem do pesquisador, como o tratamento e disponibilização dos dados das pesquisas que financiam.

#### 4.1.5. Licenças de acesso e uso

Sobre os tipos de licenças adotadas pelos repositórios, o HomoMINT, o American Type Culture Collection, o InnateDB, o NeuroMorpho, e o Neuroscience Information Framework usam Creative Commons. O Bii usa a Mozilla Public License, e o Domino o DOI (o DOI não é uma licença porém está comumente associado a elas por se tratar de uma ferramenta de identificação e legitimação), todos os outros repositórios não apresentavam informações sobre o assunto (Gráfico 6).

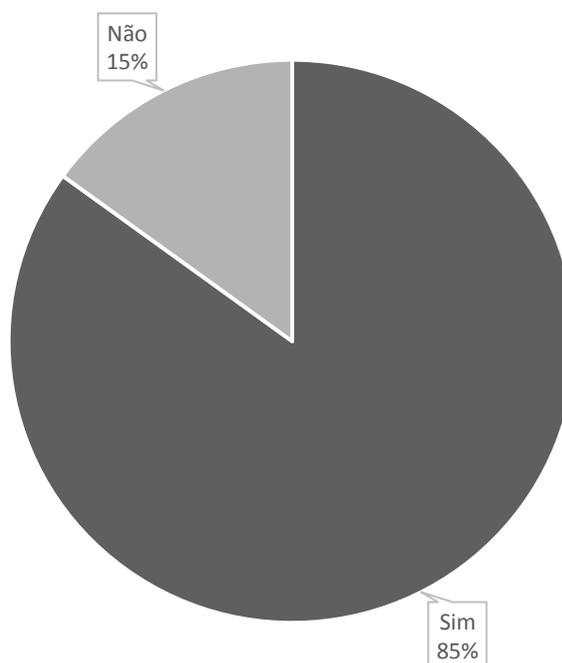
**Gráfico 6- Licenças**

Fonte: Elaboração própria

A partir do Gráfico 6 pode-se perceber que a maioria dos repositórios (13) também não informa que tipo de licença utiliza. Apenas 5 declaram utilizar a Creative commons, onde existe a possibilidade do autor abdicar de alguns de seus direitos autorais para beneficiar o acesso. Um repositório usa a Mozilla public license, feita para softwares livres e com o código aberto, é similar ao copyleft porém mais flexível ao tratar da distribuição de trabalhos derivados. Um repositório utiliza a DOI, que é um sistema de identificação de objetos na internet e por isso está comumente associado ao uso de algum copyright.

#### 4.1.6. Iniciativa de acesso aberto

A maioria dos repositórios, 85%, são iniciativas de acesso aberto (Gráfico 7). Apenas 3 repositórios, o American Type Culture Collection, Collaborative Psychiatric Epidemiology Surveys e o Danish Data Archive não são 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.

**Gráfico 7- Acesso aberto**

Fonte: Elaboração própria

Podemos lembrar do primeiro princípio do documento OECD, a abertura, é um princípio que pode ser considerado como alcançado, pois 85% dos repositórios da amostragem são de acesso aberto

#### 4.1.7. Padrão de metadados

A segunda categoria “descrição” diz respeito ao padrão de metadados. Neste aspecto foi identificada a utilização de 9 padrões de metadados, conforme mostra o Quadro 5. Nos outros repositórios não foram fornecidas informações.

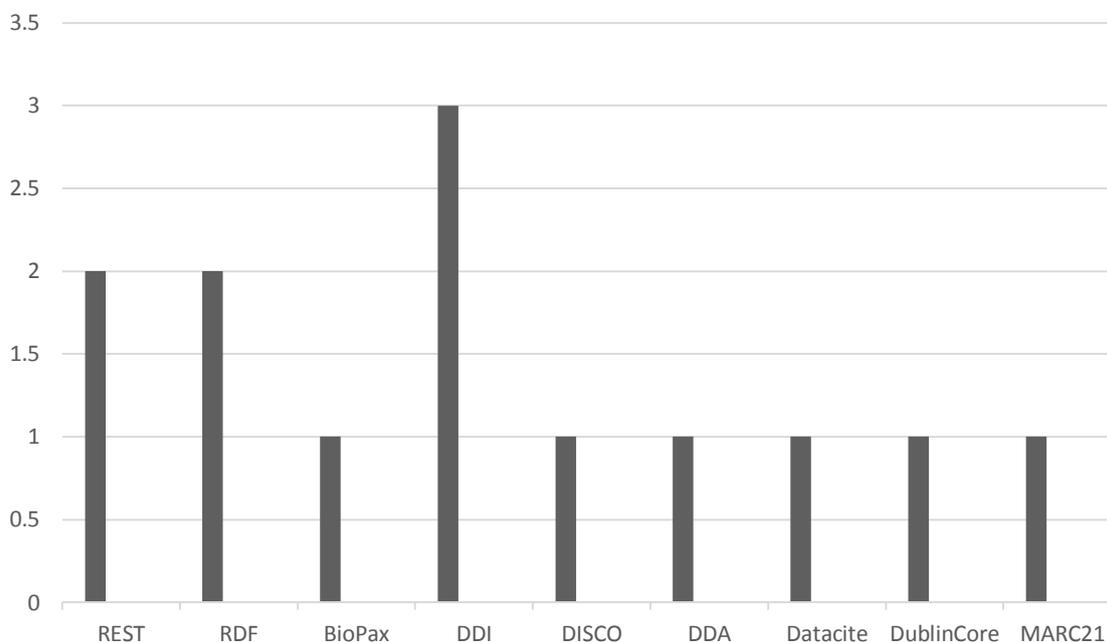
**Quadro 5- Padrão de metadados**

Repositório	Padrão
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

Alguns repositórios de dados de pesquisa se utilizam de padrões de metadados em comum, conforme o Gráfico 8 ilustra a situação.

**Gráfico 8-** Padrão de metadados



Fonte: Elaboração própria

O gráfico acima denota que o padrão mais utilizado é o DDI (com 3 repositórios), seguido do REST e do RDF empatados (com 2 repositórios cada). Todos os outros padrões encontrados são usados em apenas 1 repositório cada.

O DDI<sup>8</sup> é um padrão internacional desenvolvido pela DDI Alliance para descrever pesquisas, censos, e outras atividades de coleta de microdados, voltado especialmente para dados obtidos a partir de métodos de observação nas ciências sociais, comportamentais, econômicas e da saúde.

O RDF<sup>9</sup> é um padrão voltado para facilitar a junção de dados criando modelos simples, já o padrão REST<sup>10</sup> de metadados foca nas interações e restrições destacando dados significantes.

O BioPax<sup>11</sup> 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.

<sup>8</sup> Fonte: <http://www.ddialliance.org/>

<sup>9</sup> Fonte: <https://www.w3.org/RDF/>

<sup>10</sup> Fonte: <http://apiux.com/2013/04/09/rest-metadata-formats/>

<sup>11</sup> Fonte: <http://www.biopax.org/>

Já DISCO<sup>12</sup> é um método de integração de informação baseado no protocolo de descoberta da Microsoft para facilitar interoperação entre recursos da internet.

DDA<sup>13</sup> é o padrão próprio do Danish Data Archive, já o Datacite é o esquema padrão da organização Datacite para publicação e citação de dados de pesquisa.

O DublinCore<sup>14</sup> é um esquema de metadados que objetiva a descrição de objetos digitais a partir de um vocabulário de 15 propriedades utilizado na descrição do recurso. Por fim o MARC21<sup>15</sup>, 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.

#### 4.1.8. Descrição dos recursos

Sobre os campos de descrição onde foram definidas 3 características chave, que são:

- 1- Identificação (autoria, instituição, etc.),
- 2- Descrição dos dados e,
- 3- Permissões, sendo analisadas sua presença e extensão e destacadas anomalias.

Podemos destacar os repositório HomoMINT e o Bacterial Carbohydrate Structure DataBase com relação ao quesito identificação, disponibilizando por exemplo autoria, instituição, nome do periódico, entre outros. Ainda sobre a identificação os repositórios mentha e Hardin.MD apresentam uma seleção pobre de campos, destacando apenas informações técnicas do dado, como código do gene ou sintomas.

Ainda sobre os campos de descrição, ao se tratar das descrições dos dados vale destacar os repositórios MaizeGDB e o Human Proteinpedia por apresentarem uma ampla abrangência de campos relacionais. Por outro lado os repositórios Hardin.MD, Bii, ClinicalTrials.gov e o Danish Data Archive não disponibilizam muitas informações sobre os dados ofertados dificultando a recuperação e acesso dos dados desejados. Por fim ao se tratar das permissões apenas os repositórios American Type Culture Collection, INDEPTH Data Repository, Bii, ClinicalTrials.gov e o Collaborative Psychiatric Epidemiology Surveys apresentam campos relacionados. Vale destacar que grande parte dos repositórios disponibilizavam o link de acesso ao PubMed, demonstrando uma preocupação de relacionar os dados às publicações.

---

<sup>12</sup> Fonte: <http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14/paper/view/151>

<sup>13</sup> Fonte: <https://www.sa.dk/>

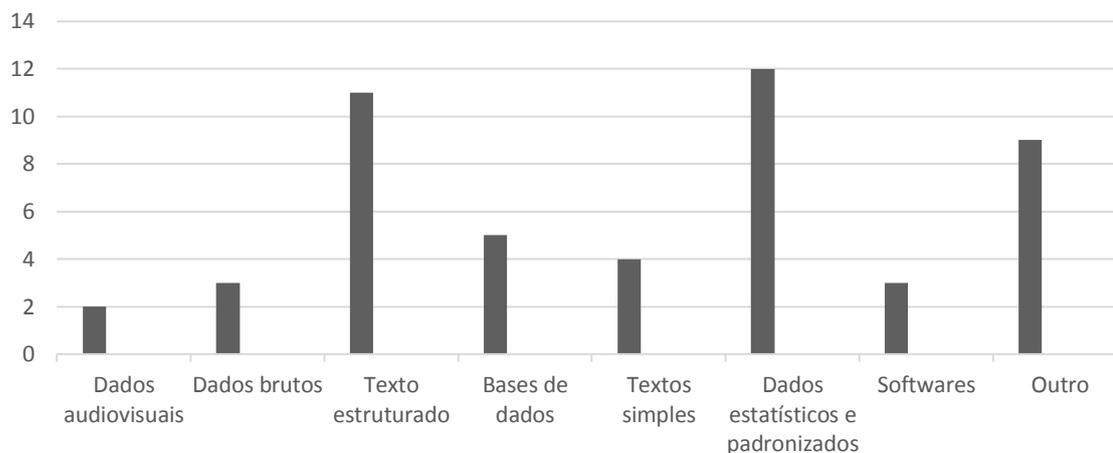
<sup>14</sup> Fonte: <http://dublincore.org/>

<sup>15</sup> Fonte: <http://www.bibliosys.com.br/bibliosys/informativo/faq/id/30/?1-O-que-e-o-MARC-21.html>

#### 4.1.9. Tipos de recurso

Quanto aos tipos de recurso foram encontrados os seguintes resultados:

- Dados audiovisuais, (Nos repositórios HomoMINT e NeuroMorpho),
- Dados brutos (HomoMINT, American Type Culture Collection e DiversityData.org),
- Texto estruturado, ou texto apresentado em formatos diferenciados, como em tópicos ou tabelas (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, apresentado de forma direta e sem o uso de artifícios visuais (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),
- Software (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).

**Gráfico 9-** Tipo de recurso

Fonte: Elaboração própria

De acordo com o Gráfico 9 acima podemos perceber que dados estatísticos e padronizados são os tipos de recursos mais comuns, com 12 repositórios. Seguido imediatamente por texto estruturado com 11 repositórios. Os tipos menos comuns são os dados audiovisuais, com 2 repositórios e os softwares com 3.

De novo usando o documento OECD O décimo primeiro princípio, eficiência, está em boas condições, a maioria dos repositórios apresenta dados diferenciados ou ao menos de forma diferenciada, o retrabalho também é evitado em parte graças a iniciativa PUBMED.

#### 4.1.10. Tamanho dos repositórios

Com relação à quantidade de itens observou-se grande disparidade. Os resultados sobre a quantidade de itens foram organizados por ordem decrescente:

- Neuroscience Information Framework- 829.679.866 itens,
- Human Proteinpedia- 7.038.972 itens.
- ArrayExpress- 1.962.932 itens,
- InnateDB- 888.753 itens,
- Mentha- 700.745 itens,
- HomoMINT- 343.277 itens,
- ClinicalTrials.gov- 206.902 itens,
- American Type Culture Collection- 74.000 itens,
- Bacterial Carbohydrate Structure DataBase- 21.904 itens,

- Domino- 15.981 itens,
- Collaborative Psychiatric Epidemiology Surveys- 6.274 itens,
- INDEPTH Data Repository- 4.239 itens,
- Bii- 133 itens.

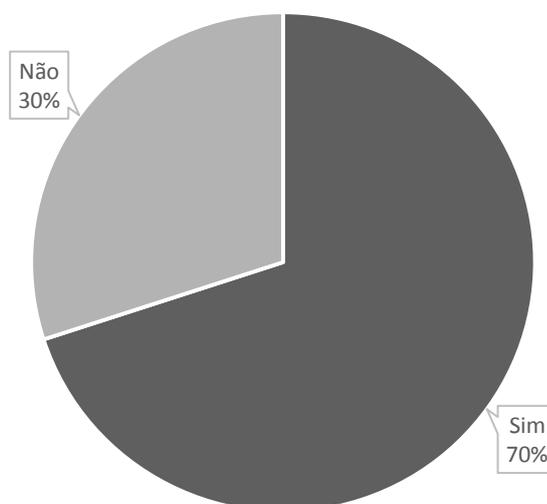
Os repositórios MaizeGDB, Hardin.MD, Canadian Epigenetics, Environment and Health Research Consortium Platform, Centers for Disease Control and Prevention, Data & Statistics, Danish Data Archive, DiversityData.org e o NeuroMorpho não informam o número de entradas em seus bancos de dados.

Percebe-se essa distância no número de itens tão discrepante devido ao escopo de cada repositórios, enquanto alguns consideram um conjunto de dados completo como um único item outros tratam são quase catálogos de moléculas e interações, como nos casos dos Neuroscience Information Framework e da Human Proteinpedia.

#### 4.1.11. Relação dos dados com seu conjunto

Quanto a informações sobre a relação do item com o seu conjunto (o conjunto pode ser entendido como conjunto de dados similares, ou de um mesmo estudo, entre outros tipos de acordo com a proposta do repositório), como destacado pelo Gráfico 10, na maioria dos repositórios (14 deles) é apresentado uma relação dos itens com o conjunto de uma forma geral. Vale lembrar que a resposta negativa não necessariamente significa que a relação não exista, apenas que não nos é informado.

Os repositórios que não apresentam informações sobre a relação são: American Type Culture Collection, Mentha, Hardin.MD, Bii, Canadian Epigenetics, Environment and Health Research Consortium Platform e o Domino.

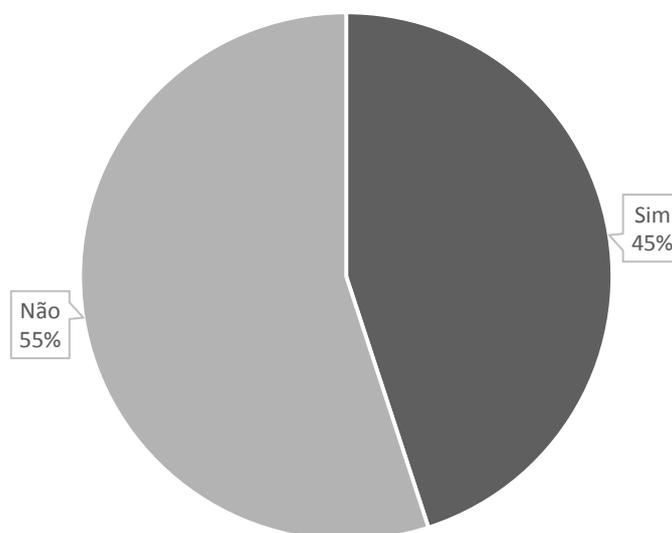
**Gráfico 10-** Relação do item com o conjunto

Fonte: Elaboração própria

#### 4.1.12. Relação do conjunto de dados da publicação

Por fim o Gráfico 11 demonstra que no quesito da relação do conjunto de dados com a publicação os repositórios encontram-se divididos, com 11 repositórios não apresentando essa relação e 10 a apresentando.

Se existe relação do conjunto com a publicação, os repositórios que apresentam essa relação são: HomoMINT, MaizeGDB, INDEPTH Data Repository, Centers for Disease Control and Prevention, Data & Statistics, ClinicalTrials.gov, Collaborative Psychiatric Epidemiology Surveys, Danish Data Archive, DiversityData.org e o Neuroscience Information Framework.

**Gráfico 11-** Relação do conjunto com a publicação

Fonte: Elaboração própria

Voltando ao documento OECD o terceiro princípio, transparência, deixa a desejar em diversos aspectos. Como observado a maioria dos repositórios não apresenta as relações dos itens com outros elementos, não referenciam a instituição do pesquisados e não oferecem diversas informações sobre os próprios repositórios.

Com relação aos outros princípios (4, 5, 6, 7, 10, 12, 13) não são disponibilizadas informações suficientes para contatações. Segue então um tabela com os pontos fortes encontrados nos repositórios em acordo com o documento OECD.

**Quadro 6-** Pontos fortes

Pontos fortes
<b>Abertura</b> , pela quantidade de repositórios de acesso aberto que disponibiliza o acesso com baixos custos
<b>Qualidade</b> , foi encontrada uma grande conformidade com os padrões de qualidade, tanto os estabelecidos pelos próprios repositórios como por outras entidades
<b>Eficiência</b> , pela comunicação com o PUBMED ou por apresentar os dados de diferentes formas, dados repetidos ou retrabalho é evitado

Fonte: Elaboração própria

Percebe-se então que temos tópicos que ainda merecem atenção e aprimoramento ao se tratar de uma forma geral os repositórios de dados, mas que em sua maioria eles já apresentam condições em estado satisfatório ou ótimo de acordo com os parâmetros propostos pelo OECD. Na próxima e última seção, Conclusões, encontraremos as considerações finais do tema.

## 5. Conclusões

A partir do método de pesquisa utilizado foram recuperados 20 repositórios certificados no RE3DATA, desses repositórios os que se encontraram em melhor estado foram três. O Neuroscience Information Framework, com uma interface amigável, com transparência de informações e fácil acesso aos conjuntos de dados. O DiversityData.org com uma navegação intuitiva e de fácil leitura dos dados e de suas relações. Por fim, o Bacterial Carbohydrate Structure DataBase apresentando a melhor transparência e disponibilizando uma extensa variedade de modos de leitura de dados.

Os repositórios que apresentaram maiores dificuldades também foram três. O Domino com um sistema de busca lenta e confusa e pouca transparência apesar da interoperabilidade com outros repositórios. O Danish Data Archive pela barreira linguística e problemas de navegação no site. Finalmente, o Bii pela confusa organização e disposição dos dados de pesquisa.

Os repositórios analisados se encontram em estados distintos, alguns já bem avançados e organizados e outros deixando a desejar, porém de uma forma geral percebe-se o avanço desses repositórios em diversos fatores. O que faz com que se essa tendência se mantiver no futuro os repositórios de dados de pesquisa serão mais utilizados e terão mais visibilidade devido a sua crescente demanda e benefícios ao processo científico de forma geral.

Em sua maioria os repositórios encontrados então são categorizados como disciplinares, são provenientes dos Estados Unidos, e são de acesso aberto. A maioria não informa sobre sistema de preservação, porém dos repositórios que informam a maior parte utiliza a curadoria manual. Não é informado na maior parte deles sobre os softwares utilizado e sobre a instituição do pesquisador e licença utilizada, porém dos que informam sobre seu licenciamento, a maior parte utiliza a creative commons.

O padrão de metadados mais utilizado é o DDI. A maior parte apresenta dados estatísticos e padronizados, e o maior dos repositórios em quantidade de itens é o Neuroscience Information Framework, sendo o Bii o menor encontrado. Sobre a relação do item com o conjunto e por fim a maioria apresenta essa relação e aproximadamente metade não apresenta relação do conjunto com a publicação.

A pesquisa se limitou aos repositórios certificados no site RE3DATA, existindo diversos outros repositórios que ainda não se encontram certificados. Ela também foi limitada aos dados fornecidos pelos próprios repositórios, que muitas vezes não são completas ou não correspondem completamente às questões levantadas.

Como recomendações para estudos futuros sugere-se a ampliação da amostra, bem como a investigação sobre políticas mandatórias de depósito de dados de pesquisa, que podem impulsionar grandemente o avanço da área ou estudar a relação das instituições e suas características e os dados de pesquisa por ela produzidos ou disponibilizados. Recomenda-se também um estudo sobre a relação dos metadados e os tipos de recursos.

## Referências bibliográficas

BAPTISTA, Ana Alice. et. al. Comunicação científica: o papel da Open Archives Initiative no contexto do Acesso Livre. **Encontros Bibli**, Porto Alegre, n. esp, 17 p. 2007. Disponível em: <<http://repositorium.sdum.uminho.pt/handle/1822/8727>>. Acesso em: 22 de dez. de 2015.

BIOMED CENTRAL. **Editorial policies**. Disponível em: <<http://www.biomedcentral.com/submissions/editorial-policies#availability+of+data+and+materials>>. Acesso em: 22 de dez. de 2015.

BRODY, Tim. et. al. Incentivizing the Open Access Research Web Publication-Archiving, Data-Archiving and Scientometrics. **CTWatch Quarterly**, v. 3, n. 3, 9 p. ago. 2007. 2007. 9 p. Disponível em: <<http://eprints.soton.ac.uk/264418/>>. Acesso em: 22 de dez. de 2015.

COSTA, Michelli Pereira da. **Características e contribuições da via verde para o acesso aberto à informação científica na américa latina**. Brasília: Universidade de Brasília, 2014. 226 p. Disponível em: <<http://repositorio.unb.br/handle/10482/15687>>. Acesso em: 22 de dez. de 2015.

COSTA, Michelli Pereira da; LEITE, Fernando César Lima. Repositórios institucionais de acesso aberto à informação científica: proposta de modelo de avaliação. **RECIIS – Rev Eletron de Comun Inf Inov Saúde**. Rio de Janeiro, v.9, n.3, 9 p. jul./set. 2015. Disponível em: <<http://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/996/pdf996>>. Acesso em: 22 de dez. de 2015.

COSTA, Sely. Abordagens, estratégias e ferramentas para o acesso aberto via periódicos e repositórios institucionais em instituições acadêmicas brasileiras. **Liinc em Revista**, Rio de Janeiro, v.4, n.2, p. 218 – 232, set. 2008. Disponível em: <<http://revista.ibict.br/liinc/index.php/liinc/article/viewFile/281/172>>. Acesso em: 22 de dez. de 2015.

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

DECLARAÇÃO BUDAPEST OPEN ACCESS INITIATIVE. **Read the Budapest Open Access Initiative**. Budapest, 2002. Disponível em: <<http://www.budapestopenaccessinitiative.org/read>>. Acesso em: 03 de fev. de 2016

DECLARAÇÃO DE BERLIM. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. Disponível em: <<http://openaccess.mpg.de/Berlin-Declaration>>. Acesso em: 03 de fev. 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.

FERREIRA, Andressa Oliveira. **Aspectos éticos envolvidos no processo de compartilhamento de dados de pesquisa**. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2012. 60 p. Disponível em: <<http://www.lume.ufrgs.br/handle/10183/69764>>. Acesso em: 22 de dez. de 2015.

GARVEY, William D. **Communication: The Essence of Science: Facilitating Information Exchange Among Librarians, Scientists, Engineers and Students**. :Elsevier, 2014. 344 p.

GIGLIA, E. Open Acces in the biomedical field: a unique opportunity for researchers (ans research itself). **European journal of physical and rehabilitation medicine- europa medicophysica**. p. 203-213, jun. 2007. Disponível em: <<http://www.minervamedica.it/en/journals/europa-medicophysica/article.php?cod=R33Y2007N02A0203>>. Acesso em: 22 de dez. de 2015.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. 4. reimpr. São Paulo: Atlas, 2011. 216 p.

GLOVER, Steven William. et. al. Open access publishing in the biomedical sciences: could funding agencies accelerate the inevitable changes?. **Health Information and Libraries Journal**, v. 23, n. 3, p. 197-202, set. 2006. Disponível em: <<http://onlinelibrary.wiley.com/doi/10.1111/j.1471-1842.2006.00657.x/abstract;jsessionid=D5E30614723147B6C34A69CCAD50E80A.f04t03>>. Acesso em: 22 de dez. de 2015.

HARNAD, Steven. Ethics of open access to biomedical research: Just a special case of ethics of open access to research. **Philosophy, ethics, and humanities in medicine**. 5 p. dez. 2007. Disponível em: <<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2228286/>>. Acesso em: 22 de dez. de 2015.

LATINDEX. **Descripción**. Acesso em: 05 de jan. de 2016. Disponível em: <<http://www.latindex.org/latindex/descripcion>>.

HURD, Julie M. The transformation of scientific communication: A model for 2020. **Journal of the american society for information science**. p. 1279-1283. dez. 2000

MATSUBAYASHI, Mamiko. et. al. Status of open access in the biomedical field in 2005Disponível em: <>. Acesso em: 22 de dez. de 2015. **Journal of the Medical Library Association**. p. 4-11. jan. 2009. Disponível em: <<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605039/>>. Acesso em: 22 de dez. de 2015.

MEDEIROS, Jackson da Silva; CAREGNATO, Sônia Elisa. Compartilhamento de dados e e-Science: explorando um novo conceito para a comunicação científica. **Liinc em Revista**, Rio de Janeiro, v.8, n.2, p. 311-322, set. 2012. Disponível em <<http://revista.ibict.br/liinc/index.php/liinc/article/view/488>>. Acesso em: 22 de dez. de 2015.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, OECD. **OECD Principles and Guidelines for Access to Research Data from Public Funding**. Paris: OECD PUBLICATION, 2007. 22 p. Disponível em: <<https://www.oecd.org/sti/sci-tech/38500813.pdf>>. Acesso em 15 de abr. de 2016.

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

PUBLIC LIBRARY OF SCIENCE, PLoS. **Data Availability**. 2014. Disponível em: <<http://journals.plos.org/plosone/s/data-availability>>. Acesso em: 05 de jan. de 2016.

PUBMEDCENTRAL, PMC. **PMC Overview**. 2000. Acesso em: 05 de jan. de 2016. Disponível em: <<http://www.ncbi.nlm.nih.gov/pmc/about/intro/>>.

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

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 2015.

ROSENDAAL, Hans E.; GEURTS, Peter A. Th. M. **Forces and functions in scientific communication**: an analysis of their interplay. 1998. 22 p. Disponível em: <<http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html>>. Acesso em: 22 de dez. de 2015.

SANS VALERO, Javier; VEIGA DE CABO, Jorge. A iniciativa Open Access no acesso à informação técnico-científica nas Ciências da Saúde. **RECIIS – R. Eletr. de Com. Inf. Inov. Saúde**. Rio de Janeiro, 2007, v. 1, n. 1, p. 19-26. Disponível em: <<http://eprints.rclis.org/12212/>>. Acesso em: 22 de dez. de 2015.

SARMENTO, Fernanda. et. al. **Algumas considerações sobre as principais declarações que suportam o movimento Acesso Livre**. World Congress on Health Information and Libraries. Salvador: 2005. Disponível em: <<http://hdl.handle.net/10760/8512>>. Acesso em: 22 de dez. de 2015.

SCIENTIFIC ELETRÔNIC LIBRARY ONLINE, SciELO Saúde Pública. **Critérios, política e procedimentos para a admissão e a permanência de periódicos científicos na Coleção SciELO Brasil**. São Paulo, 2014. Acesso em: 05 de jan. de 2016. Disponível em: <[http://www.scielo.br/avaliacao/20141003NovosCriterios\\_SciELO\\_Brasil.pdf](http://www.scielo.br/avaliacao/20141003NovosCriterios_SciELO_Brasil.pdf)>.

SCIENTIFIC ELETRÔNIC LIBRARY ONLINE, SciELO Saúde Pública. **Sobre esse site**. Disponível em: <<http://www.scielosp.org/?lng=pt>>. Acesso em: 22 de dez. de 2015.

SUBER, Peter. Open access policy options for funding agencies and universities. **SPARC Open Access Newsletter**. 26 p. fev, 2009. Disponível em: <<https://dash.harvard.edu/handle/1/4322589>>. Acesso em: 22 de dez. de 2015.

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

## Apêndice A

**Quadro 7- Dimensões da análise: Identificação**

Repositórios	Tipo de repositório	País	Instituição
HomoMINT	Disciplinar	Itália	Universita degli Studi di Roma 'Tor Vergata'
American Type Culture Collection	Disciplinar	Estados Unidos da América	LGC Standards
MaizeGDB	Institucional	Estados Unidos da América	Iowa State University, Department of Agronomy
ArrayExpress	Governamental	Reino Unido	EMBL-EBI, Functional Genomics Group
Bacterial Carbohydrate Structure DataBase	Governamental	Rússia	The President of the Russian Federation Grants Council
InnateDB	Institucional	Canadá	Simon Fraser University, Fiona Brinkman Laboratory
mentha	Institucional	Itália	Universita degli Studi di Roma 'Tor Vergata'
INDEPTH Data Repository	Disciplinar	Gana	INDEPTH Network
Human Proteinpedia	Disciplinar	Índia	Institute of Bioinformatics
Hardin.MD	Institucional	Estados Unidos da América	The University of Iowa Libraries
Bii	Institucional	Alemanha	Leibniz Institute of Plant Biochemistry,
Canadian Epigenetics	Disciplinar	Canadá	Canadian Institutes of Health Research
Centers for Disease Control and Prevention	Disciplinar	Estados Unidos da América	Centers for Disease Control and Prevention
ClinicalTrials.gov	Disciplinar	Estados Unidos da América	HHS.gov
Collaborative Psychiatric Epidemiology Surveys	Disciplinar	Estados Unidos da América	National Institute of Mental Health
Danish Data Archive	Institucional	Dinamarca	Danish Data Archives
DiversityData.org	Disciplinar	Estados Unidos da América	Brandeis University,
domino	Disciplinar	Itália	Universita degli Studi di Roma 'Tor Vergata'
NeuroMorpho	Institucional	Estados Unidos da América	George Mason University
Neuroscience Information Framework	Governamental	Estados Unidos da América	Neuroscience Information Framework

Fonte: Elaboração própria

## Apêndice B

Quadro 8- Dimensões da análise: Armazenamento

Repositórios	Software	Sistema de Preservação	URL persistente	Instituição do pesquisador	Licença	Acesso aberto
HomoMINT	MINT Viewer			N	Creative Commons	S
American Type Culture Collection				N	Creative Commons	N
MaizeGDB		Curadoria comunitária		N		S
ArrayExpress				N		S
Bacterial Carbohydrate Structure DataBase	MySQL			S		S
InnateDB	Cerebral program (Java)	Curadoria manual		N	Creative Commons	S
mentha	REST	Curadoria manual		N		S
INDEPTH Data Repository				S		S
Human Proteinpedia				N		S
Hardin.MD				S		S
Bii	BiolnvIndex		Sim	N	Mozilla Public License	S
Canadian Epigenetics	OpenCEMT			N		S
Centers for Disease Control and Prevention				N		S
ClinicalTrials.gov				S		S
Collaborative Psychiatric Epidemiology Surveys				S		N
Danish Data Archive				S		N
DiversityData.org				N		S
domino				N	<b>DOI</b>	S
NeuroMorpho		Curadoria centralizada		N	Creative Commons	S
Neuroscience Information Framework				N	Creative Commons	S

Fonte: Elaboração própria

## Apêndice C

**Quadro 9- Dimensões da análise: Descrição**

Repositórios	Metadados (padrão)	Campos de descrição
HomoMINT		Autoria, periódico, pubmed, identificação dos dados, interação dos dados
American Type Culture Collection		Número, classificação, cepa, formato, segurança, preços
MaizeGDB		Sumário rápido, visão geral, anotações, coordenadas do cromossomo, coordenadas no mapa, cepas próximas, alelo/variação/polimorfismo, informação
ArrayExpress	REST	Status, organismo, amostras, ordem, protocolos, descrição, tipos de experimento, contatos, miame, arquivos
Bacterial Carbohydrate Structure DataBase	RDF	Autoria, título, número, nome, ID, conexões, grupo taxonômico, pubmed, publicação, periódico, editora, correspondente, instituição
InnateDB	BioPax	ID do gene Ensembl, organismo, cromossomos, símbolo do gene, interações validadas experimentalmente, interações previstas pela orthology
mentha		Nome do gene, organismo, uniprot, nome, nome alternativo, orthologs
INDEPTH Data Repository	DDI, RDF	ID de referência, ano, país, produtores, financiadores, coleções, metadata, descrição do estudo, dicionário de dados, materiais relacionados, citações
Human Proteinpedia		Tipo de experimento, descrição curta, descrição do experimento, nome do principal pesquisador, título, e-mail, endereço, país,
Hardin.MD		Figuras, sintomas
Bii		Créditos, contato e feedback, licença, financiadores
Canadian Epigenetics		Codônimo, ID anônimo, provedor da biomatéria, tipo da biomatéria, tipo de célula, descrição, doença, idade do doador, etnia do doador
Centers for Disease Control and Prevention		Informações básicas, pesquisa, recomendações, dados e estatísticas, artigos, materiais grátis
ClinicalTrials.gov	REST	Propósito, elegibilidade, contatos e localidades, mais informações
Collaborative Psychiatric Epidemiology Surveys	DDI, Dublin Core, MARC21	Número, título, tamanho, documentação, análise online
Danish Data Archive	DDI, DDA, Datacite	Ano de início, ano final, série de arquivo, nomes oficiais, arquivo criador
DiversityData.org		Ano, região, notas e fontes, valor, escala
domino		Proteína, gene, organismo, características sequenciais, estado
NeuroMorpho		Detalhes do neurônio selecionado, artigo de referência, medidas
Neuroscience Information Framework	DISCO	Título, agência, descrição, quantidade da recompensa, data postado, data de fechamento, duração

Fonte: Elaboração própria

## Apêndice D

**Quadro 10- Dimensões da análise: Apresentação dos itens**

Repositórios	Tipo de recurso	Quantidade de itens	Relação do item com o conjunto	Relação do conjunto com a publicação
HomoMINT	Dados audiovisuais, Dados brutos, Texto estruturado	343277	S	S
American Type Culture Collection	Dados brutos, Base de dados	74000	N	N
MaizeGDB	Texto simples, Dados estatísticos e padronizados, Base de dados, Texto estruturado, Outro		S	S
ArrayExpress	Dados estatísticos e padronizados, Base de dados	1962932	S	N
Bacterial Carbohydrate Structure DataBase	Dados estatísticos e padronizados, Texto estruturado	21904	S	N
InnateDB	Texto simples, Dados estatísticos e padronizados, Softwares, Outro	888753	S	N
mentha	Texto simples, Dados estatísticos e padronizados, Softwares, Texto estruturado	700745	N	N
INDEPTH Data Repository	Dados estatísticos e padronizados, Base de dados, Outro	4239	S	S
Human Proteinpedia	Texto simples, Texto estruturado, Outro	7038972	S	N
Hardin.MD	Base de dados, Outro		N	N
Bii	Dados estatísticos e padronizados, Texto estruturado	133	N	N
Canadian Epigenetics	Dados estatísticos e padronizados, Texto estruturado		N	N
Centers for Disease Control and Prevention	Dados audiovisuais, Dados estatísticos e padronizados		S	S
ClinicalTrials.gov	Texto estruturado, Outro	206902	S	S
Collaborative Psychiatric Epidemiology Surveys	Dados estatísticos e padronizados, Outro	6274	S	S
Danish Data Archive	Dados estatísticos e padronizados		S	S
DiversityData.org	Dados brutos, Outro		S	S
domino	Texto estruturado, Outro	15981	N	N
NeuroMorpho	Dados audiovisuais, Texto estruturado		S	N
Neuroscience Information Framework	Dados estatísticos e padronizados, Base de dados, Softwares, Texto estruturado	829679866	S	S

Fonte: Elaboração própria