cadernos BAD

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

Repositórios de dados de pesquisa no mundo

Michelli Costa

Universidade de Caxias do Sul (UCS) e Universidade de Brasília (UnB)

michelli@unb.br

Tiago Braga

Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e Universidade de Brasília (UnB)

tiagobraga@ibict.br

Resumo

O estudo analisa o desenvolvimento dos repositórios de dados no mundo, a partir de uma abordagem qualitativa e quantitativa. A pesquisa utilizou como fonte de informação os registros dos repositórios disponíveis no diretório re3data.org. Foram definidas categorias e características de análise e os dados obtidos foram utilizados para a identificação do perfil dos repositórios catalogados. Como resultados do estudo observou-se que a área de ciências da vida é responsável pela maioria dos repositórios, sendo que estes estão usualmente hospedados nos países da América do Norte e Europa. Do mesmo modo, constatou-se que a maioria dos repositórios de dados de pesquisa são temáticos, focados no armazenamento de textos e imagens, com pouca representatividade no que diz respeito a certificação e padrões de metadados. No entanto, foi possível observar uma tendência de desenvolvimento de repositórios de dados baseados em três tecnologias dominantes. De forma geral, a maioria dos repositórios de dados distribuem seus conteúdos sob uma licença *Copy Right*, embora seus dados estejam caracterizados como de acesso aberto. Ao final dessa pesquisa conclui-se que para atingir os objetivos da ciência aberta para os dados de pesquisa ainda há avanços a serem alcançados.

Palavras-chave: repositórios de dados, dados de pesquisa, ciência aberta.

Repositories of research data in the world

Abstract

This paper presents a qualitative and quantitative approach aiming to analyse how the data repositories around the world are being developed. To do this, the directory

re3data.org was assessed so the repository could be categorized. Some categories and characteristics were defined and the collected data was the source to catalogue the repositories` profile. As result, the research identified that majority of repositories are linked to Life Science and is hosted by North American and European countries. The most part of repositories is theme based, focused in text and image storage, without much certification and metadata standardization concerns and based in three main technologies. As a final result, the analyses showed that most repositories has Copy Rights, but are also described as open access. As a conclusion, this paper presents that some advances are needed to achieve the research data availability, an essential element to the Open Science.

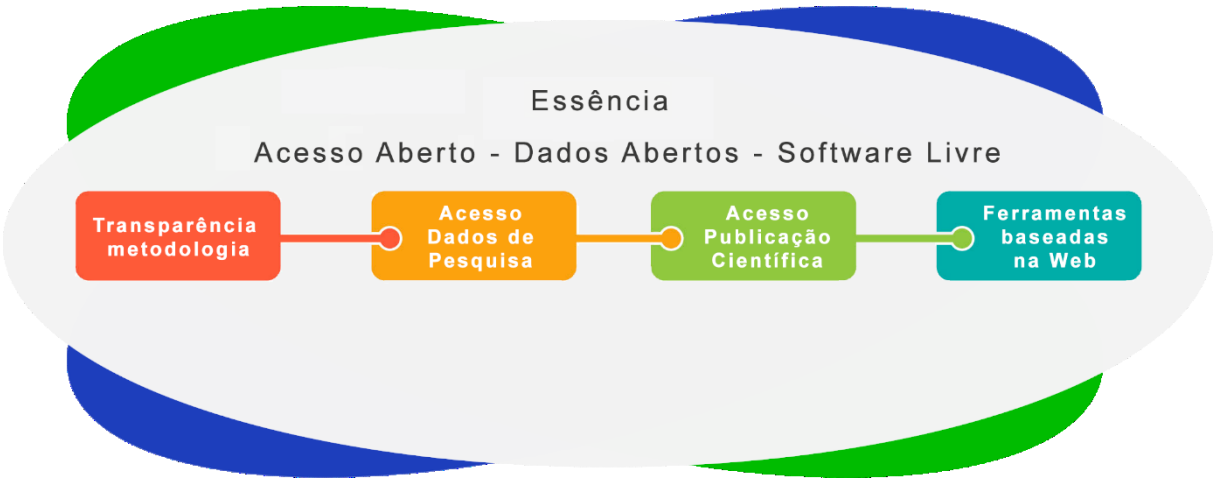
Keywords: data repositories, research data, open science.

Introdução

A emergência de novas formas de comunicação científica, intensificou a demanda por compartilhamento e reuso de dados de pesquisa, um tema relevante e contemporâneo nas discussões acerca da ciência aberta. Segundo Boulton (2013), os princípios que sustentam a ideia da ciência aberta são os mesmos que serviram de base para as revoluções científicas do século XVIII e XIX. Isto porque, em ambos os contextos a motivação seria tornar a ciência pública. O autor cita como gênese destas ideias o pioneirismo de Henry Oldenburg, que solicitou que suas cartas, enviadas à sede da associação científica Royal Society, fossem publicadas com o objetivo de atingir volumoso número de interessados sobre o tema. Oldenburg também solicitou que a publicação das cartas fosse feita em sua língua vernácula, e que fossem publicadas suas evidências para comprovar suas teorias. Neste caso, entendem-se as evidências como os dados coletados pela pesquisa e utilizados para sustentação teórica de sua argumentação. Conforme o autor, a prática de tornar públicas as evidências que sustentam uma teoria favorecem o princípio científico da refutação e do que ele chamou de *scientific self-correction*.

Os dados de pesquisa também são elementos essenciais para o cumprimento dos quatro objetivos da ciência aberta apontados por Gezelter (2009). O primeiro deles trata da transparência da metodologia, observação e coleta dos dados. O segundo objetivo pontua a disponibilidade pública dos dados de pesquisa e permissão para sua reutilização. O terceiro diz respeito a disponibilizar abertamente as publicações científicas. O quarto objetivo evidencia a necessidade do uso de ferramentas baseadas na web com vistas a facilitar a colaboração científica (Figura 1). De acordo com o autor, tais objetivos representam a essência dos projetos relacionados ao acesso aberto, dados abertos e softwares livres.

FIGURA 1: OBJETIVOS DA CIÊNCIA ABERTA



Fonte: Elaboração própria

Segundo definição apresentada pela OECD (2004) dados de pesquisa são registros de fatos usados como fonte primária da pesquisa e que, geralmente, são usados na comunidade científica como necessários para validar os resultados de uma pesquisa. De acordo com Walport e Brest (2011), os repositórios de dados de pesquisa têm sido apontados como uma estratégia eficiente para sua organização, preservação e compartilhamento. No entanto, destacam que estes sistemas necessitam estar bem estabelecidos e dispor de ferramentas que sejam capazes de descrever e divulgar os dados de pesquisa de modo a promover seu amplo acesso e reutilização.

São evidentes as discussões sobre a relevância dos sistemas para o melhoramento dos processos relacionados a comunicação científica e o próprio avanço da ciência, mas estudos sobre o desenvolvimento dos repositórios em nível global ainda são incipientes. Portanto, o presente estudo teve por objetivo analisar o desenvolvimento dos repositórios de dados no mundo.

Metodologia

A pesquisa é de natureza mista por envolver técnicas e métodos de análise dos dados dentro da perspectiva qualitativa e quantitativa. O universo pesquisado foi limitado aos repositórios cadastrados no diretório re3data.org (Registry of Research Data Repositories). O diretório se apresenta como um registro global de repositórios de dados de pesquisa e cobre diversas áreas do conhecimento. Os dados sobre os sistemas foram coletados no diretório no primeiro semestre de 2016, período em que haviam 1558 repositórios registrados.

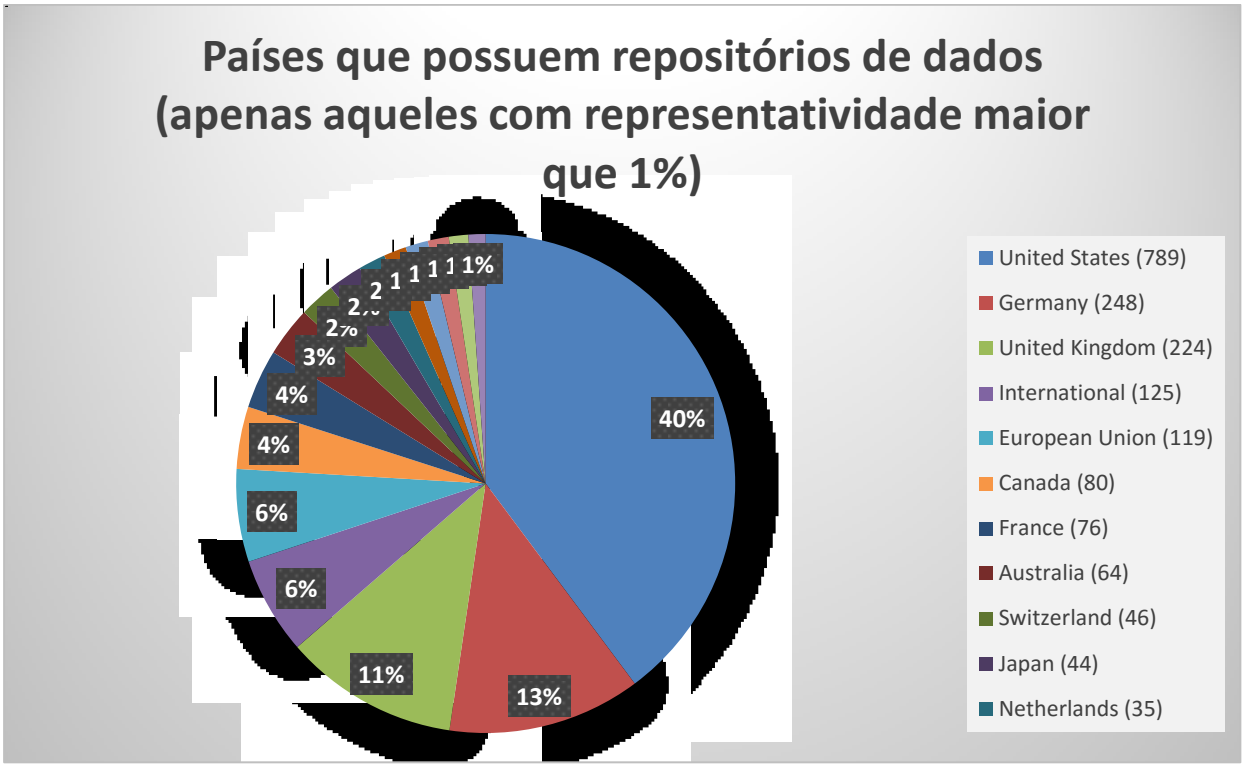
Para a coleta dos dados foram utilizados os sistemas de busca e de descoberta do diretório, que permite, entre outras funcionalidades, a organização dos dados segundo

critérios previamente determinados. Os resultados obtidos foram sistematizados em quatro categorias: abrangência geográfica, abrangência temática, características dos sistemas e promoção da ciência aberta.

Abrangência geográfica dos repositórios de dados de pesquisa

Na categoria abrangência geográfica foram identificadas e organizadas informações acerca dos países pelos quais os repositórios estão registrados. O país com maior quantidade de repositórios foram os Estados Unidos da América, de onde originam-se 40% dos repositórios de dados do mundo. A Alemanha e o Reino Unido também tiveram presença relevante no universo, representando 13% e 11% respectivamente. Além dos países da América do Norte e da Europa, apenas a Austrália, o Japão e a Índia obtiveram representação entre os países com pelo menos 1% do conjunto total de repositórios (Figura 2). Com isto, observou-se que nenhum país da América Latina e da África apresentaram expressão em quantidade de repositórios de dados de pesquisa.

FIGURA 2: ABRANGÊNCIA GEOGRÁFICA DOS DEPOSITÓRIOS DE DADOS DE PESQUISA

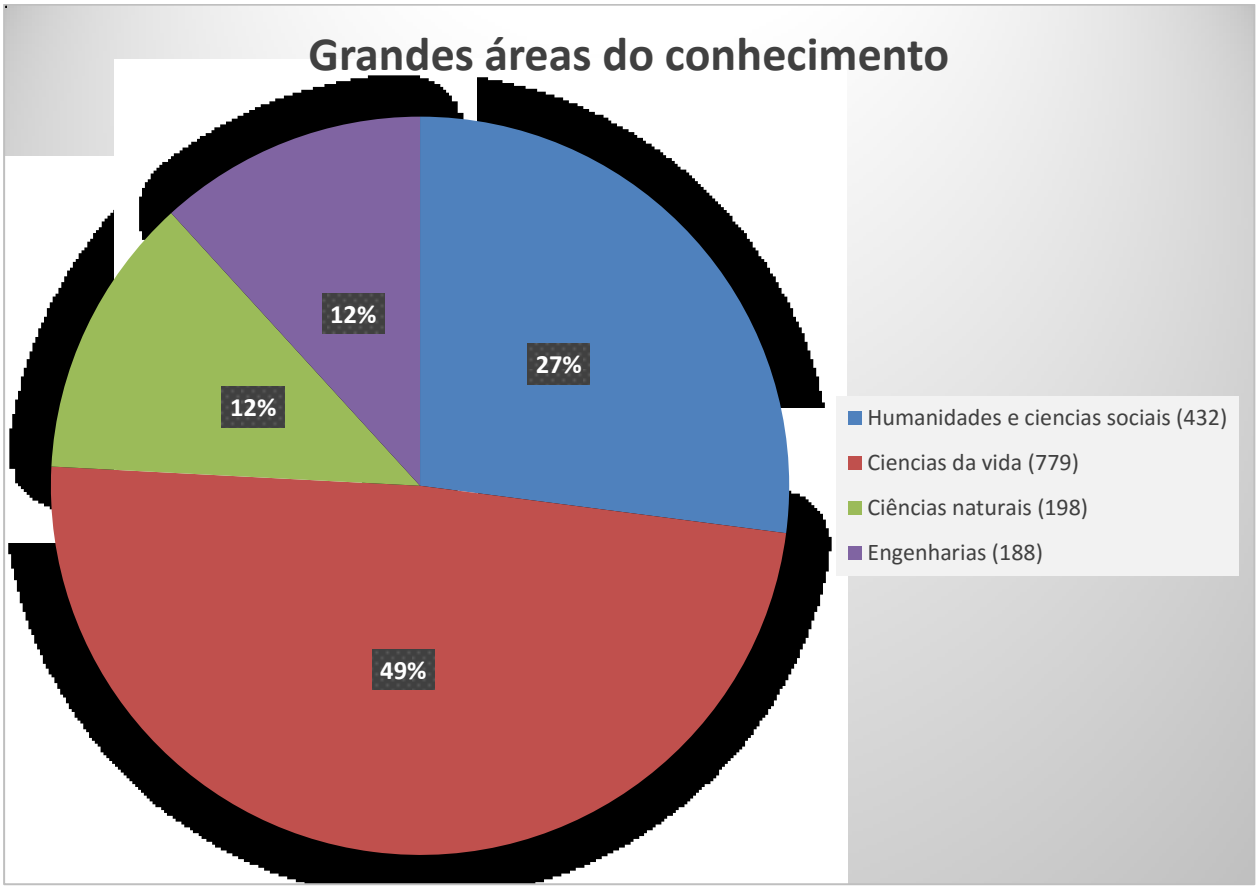


Fonte: Elaboração própria

Abrangência temática dos repositórios de dados de pesquisa

A segunda variável tratou sobre a cobertura temática dos repositórios analisados. Nesse aspecto, foi identificado que aproximadamente metade dos repositórios são relacionados às ciências da vida e a outra metade dos repositórios, dividem-se entre ciências naturais (12%), engenharias (12%), ciências sociais e humanidades (27%) (Figura 3).

FIGURA 3: ÁREAS DO CONHECIMENTO DOS REPOSITÓRIOS DE DADOS DE PESQUISA



Fonte: Elaboração própria

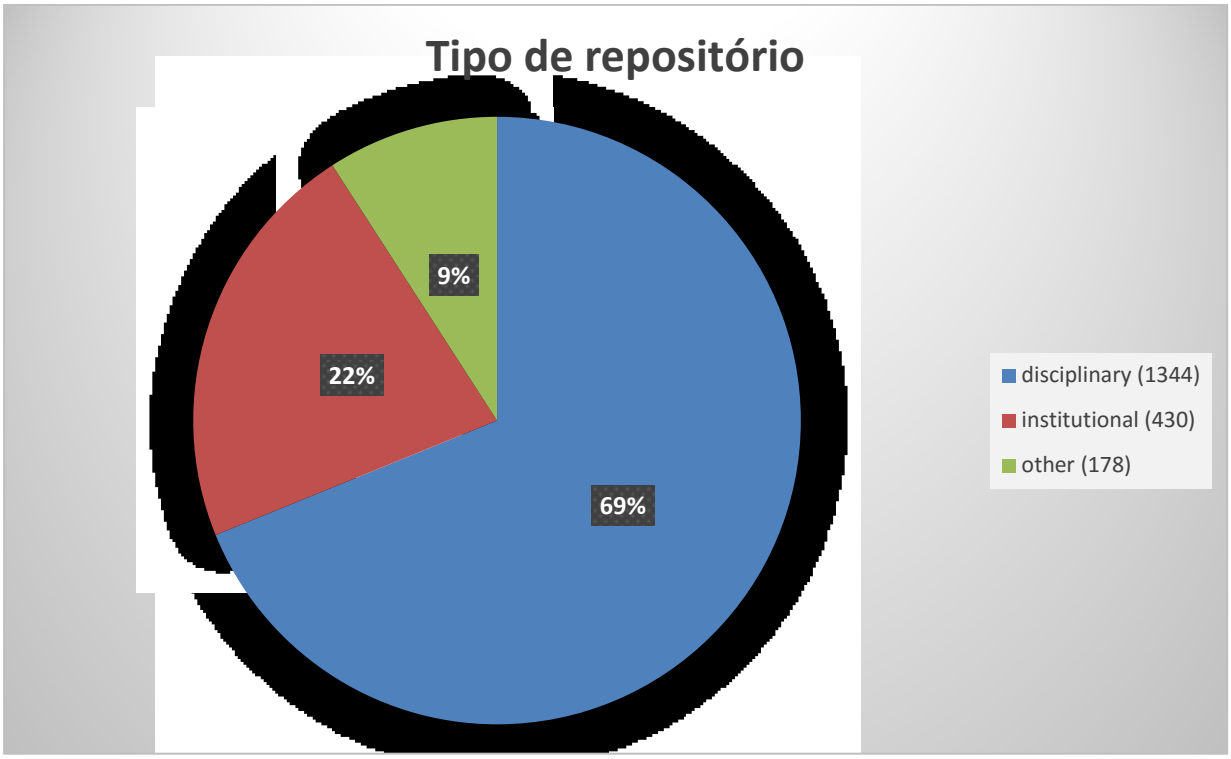
Características dos repositórios de dados de pesquisa

Na terceira categoria de análise foram exploradas cinco características dos sistemas, a saber: tipo de repositório, tipo de recursos armazenados, certificados dos repositórios, software dos repositórios e metadados.

Tipo de repositório

De forma geral, identificou-se que 69% dos repositórios de dados são temáticos, ou seja, são limitados por área do conhecimento e não por instituições (Figura 4).

FIGURA 4: TIPOS DE REPOSITÓRIOS DE DADOS DE PESQUISA



Fonte: Elaboração própria

Tipos de recursos armazenados

Dentre os documentos que os sistemas armazenam e disponibilizam, destacam-se quatro tipos: os dados científicos em formatos estatísticos (14%), documentos de texto (*office*) padrão (13%), imagens (11%) e textos simples (11%). Os outros tipos representam menos de 10% do total de documentos, inclusive os dados brutos, que são recorrentemente citados como necessários para a reutilização dos dados no contexto da ciência colaborativa (Figura 5).

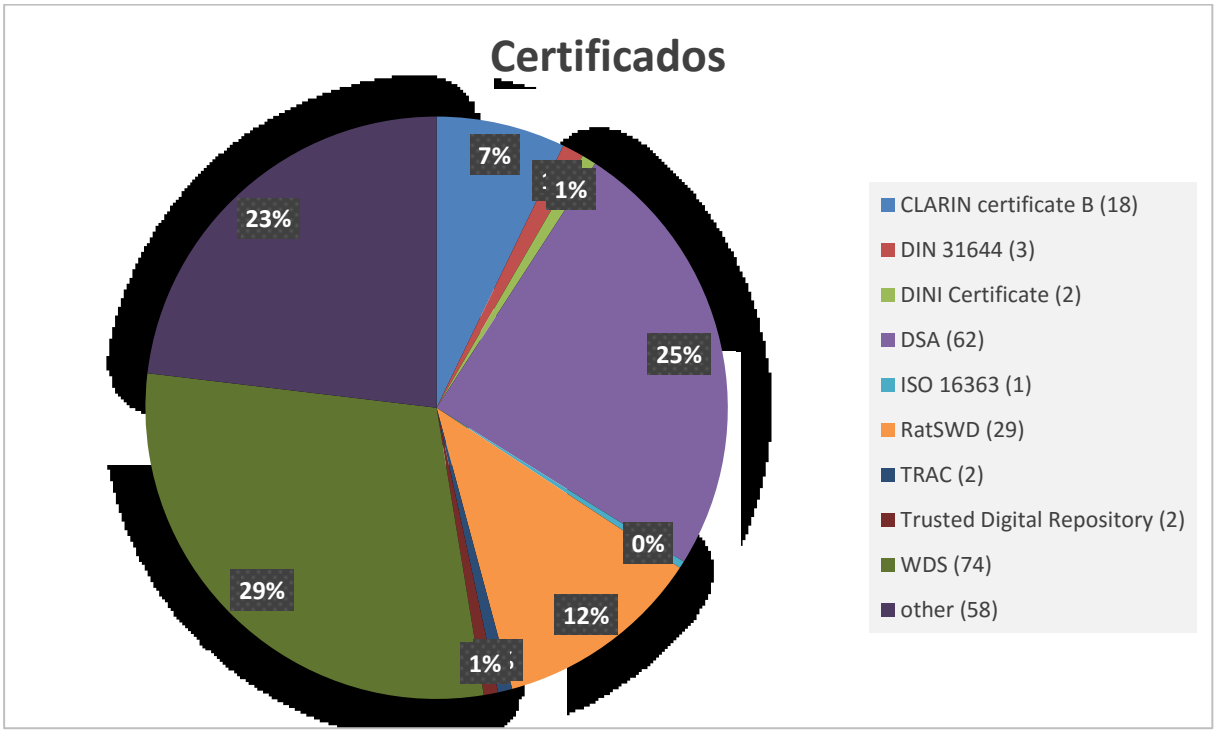
FIGURA 5: TIPOS DE RECURSOS ARMAZENADOS NOS REPOSITÓRIOS DE DADOS DE PESQUISA

Fonte: Elaboração própria

Certificados dos sistemas

Cerca de 17% dos repositórios apresentavam algum tipo de certificado de sistema. Dentre eles, destacaram-se o World Data System (WDS), em 29%; o Data Seal of Approval (DSA), em 25%; e o German Data Forum (RatSWD), em 12% destes repositórios que possuem certificados. O conjunto total de certificados utilizados pelos repositórios é apresentado na Figura 6.

FIGURA 6: CERTIFICADOS APRESENTADOS PELOS REPOSITÓRIOS DE DADOS DE PESQUISA



Fonte: Elaboração própria

O WDS é uma iniciativa do International Council for Science (ICSU) e foi criado em 2008. Seu objetivo é promover a governança de longo prazo e acesso universal e equitativo a dados científicos com garantia de qualidade, além de serviços de dados, produtos e informação.

O DSA é um selo desenvolvido pela Data Archiving and Network Services (DANS), um instituto criado em 2009 pela Academia Real de Arte e Ciência dos Países Baixos (KNAW) com apoio da Organização para Pesquisa Científica dos Países Baixos (NWO). Seu objetivo é resguardar dados, garantindo sua qualidade, bem como garantir o gerenciamento confiável dos dados para uso futuro sem a necessidade de implementar novos padrões ou onerar sua utilização.

O RatSWD é um conselho independente de diversos pesquisadores de universidades, institutos de pesquisa ligados a universidades e independentes e produtores de dados da

Alemanha. Foi criado em 2004 pelo Ministério Alemão de Educação e Pesquisa e seu objetivo é melhorar a pesquisa em infraestrutura de dados e sua competitividade internacional.

Ao se analisar os três certificados mais utilizados é possível perceber que eles possuem características bem peculiares, apesar de algumas similaridades. Do ponto de vista de abrangência, o RatSWD certifica apenas repositórios que possuem participação Alemã em sua constituição. Já o DSA e o WDS certificam repositórios independente dos países em que estes foram criados. Outra peculiaridade pode ser identificada ao se analisar o processo de criação destes certificados. Enquanto o WDS foi criado a partir de um consórcio internacional, o DSA e o RatSWD possuem sua criação baseada nos Países Baixos e na Alemanha respectivamente. O DSA foi fomentado por entidades de governo e o RatSWD, apesar de possuir respaldo do ministério alemão para Educação e Pesquisa também possui instituições privadas ligadas a sua criação.

Por fim, pode-se analisar os três certificados quanto às temáticas dos repositórios depositados. Enquanto 92% dos repositórios que utilizam o WDS armazenam dados relacionados às ciências naturais ou ciências exatas e da terra o DSA certifica 94% de repositórios cujo foco é em ciências humanas, 45% em Letras e Linguística e 29% em ciências naturais. Já os repositórios certificados pelo RatSWD possuem seu foco principalmente em ciências humanas, 72%, sociais aplicadas, 59%, e saúde, 21%.

Software

As informações acerca do *software* dos repositórios de dados de pesquisa cadastrados no re3data.org foram consideradas as menos abrangentes obtidas nesse estudo. Apenas 12% dos registros apresentaram alguma indicação a respeito dos programas computacionais utilizados para gerenciar a base de dados de seus sistemas. Parte dos sistemas que não apresentaram informações relacionadas ao software (22% do total), indicaram que não haviam opções pertinentes sobre o tema no momento do registro e portanto, responderam utilizar "outro" software entre os listados.

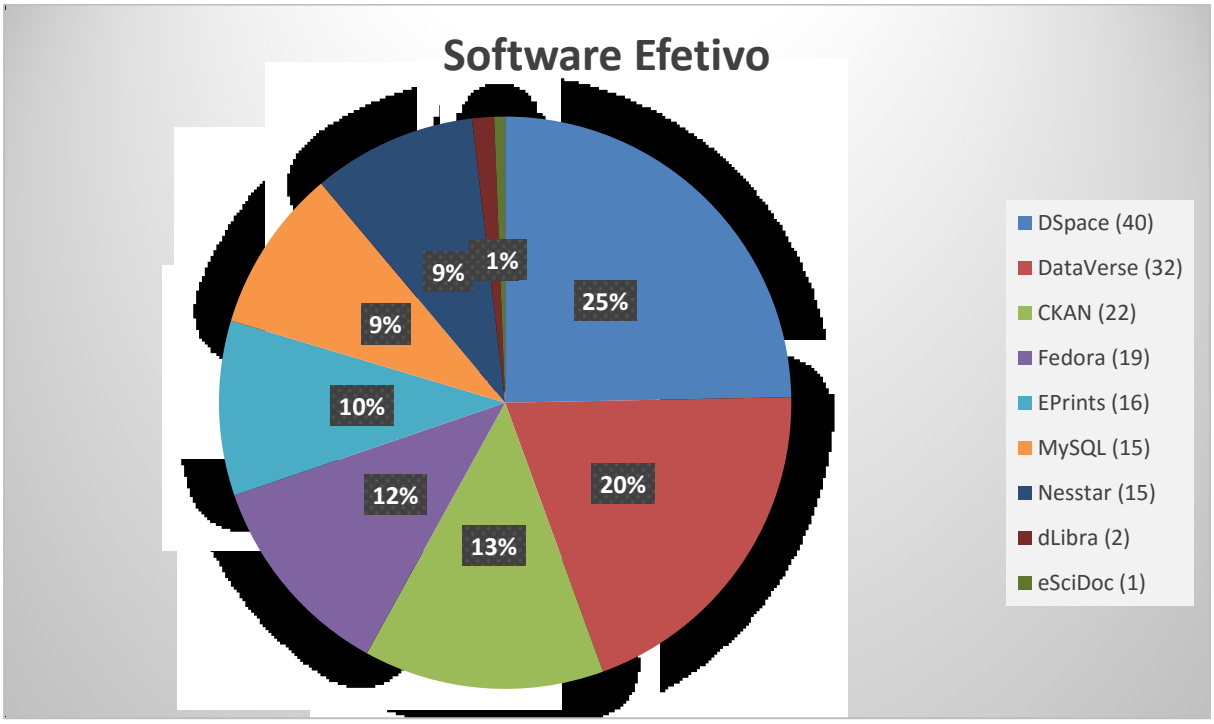
Dentre os repositórios que indicaram o software utilizado, a maioria (25%) está relacionada ao uso do DSpace (Figura 7). O software DSpace permite o acesso a todos os tipos de conteúdo digital e possui sua utilização focada em academia e organizações comerciais e sem fins lucrativos. Ele é gratuito e de código aberto e pode ser facilmente adaptado para atender as especificidades do repositório. Ele é mantido pela DuraSpace, uma organização sem fins lucrativos.

O segundo software mais utilizado para criação de repositórios de dados de pesquisa identificado pelo estudo foi o DataVerse, que está presente em 20% dos sistemas que apresentaram informação acerca do software. O DataVerse é um software de código aberto voltado para preservação, citação, exploração e análise de dados de pesquisa. Foi idealizado

através da parceria entre o Instituto para Ciência Social Quantitativa e a Universidade de Harvard. Além do software essa parceria permitiu a criação de um programa de suporte na utilização do sistema.

O software CKAN foi identificado como o terceiro mais utilizado entre os sistemas analisados, estando presente em 13% dos repositórios de dados de pesquisa. O CKAN é mantido pela Open Knowledge Foundation e possui o propósito de compartilhar, publicar, encontrar e utilizar dados. Sua utilização é focada em governos e iniciativa privada.

FIGURA 7: SOFTWARE UTILIZADOS PELOS REPOSITÓRIOS DE DADOS DE PESQUISA



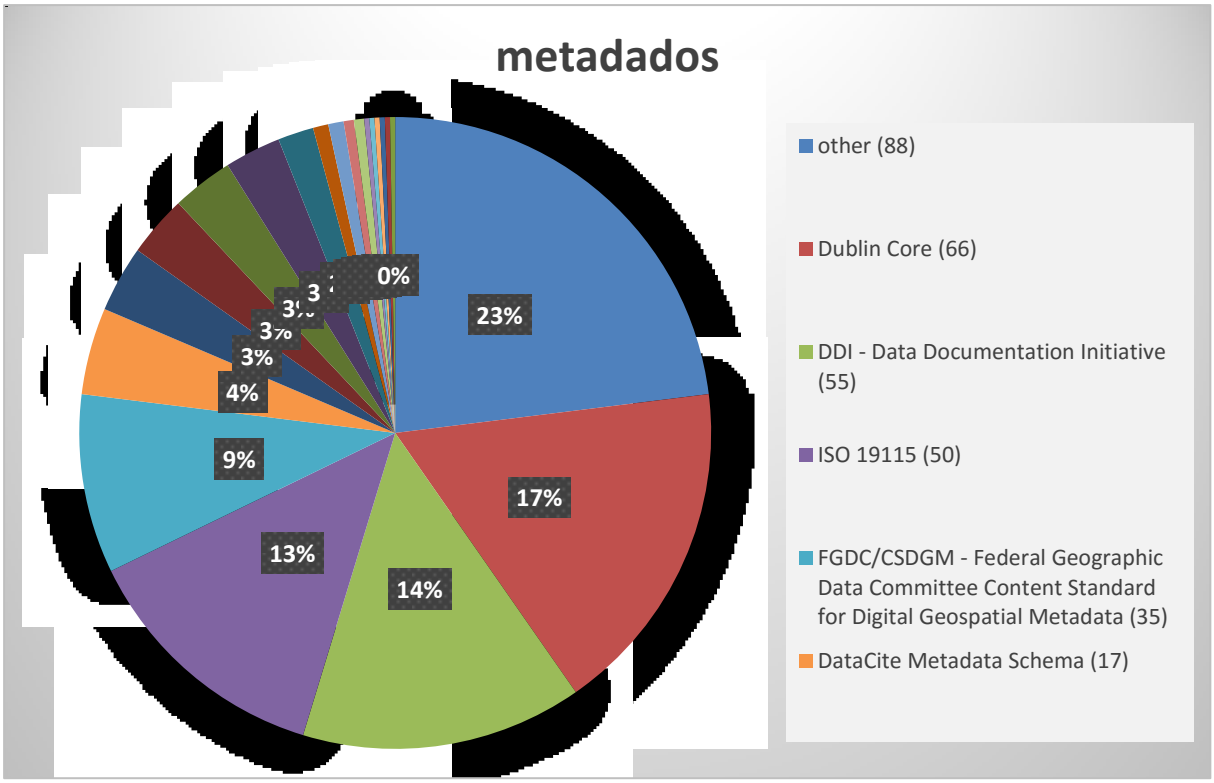
Fonte: Elaboração própria

De forma geral, o software realiza basicamente a mesma função de gestão de dados de pesquisa. Enquanto o DSpace se coloca claramente focado no ambiente acadêmico, CKAN e DataVersa não enfatizam este uso de forma tão ativa. No que se refere à mantenedora do software, tanto o DSpace quanto o CKAN são mantidos por instituições sem fins lucrativos, enquanto o DataVersa possui uma grande instituição mantenedora, a Universidade de Harvard. Uma das grandes características do DSpace é a possibilidade de adaptação do sistema para as necessidades do usuário e sua relação com uma grande comunidade de usuários e desenvolvedores.

Metadados dos repositórios

No que diz respeito à característica metadados, foi possível compreender, de forma não-abrangente, os padrões utilizados pelos repositórios. Do total de repositórios disponíveis no diretório re3data, apenas 24% deles apresentam informações relativas aos metadados utilizados para a descrição dos seus recursos. Destes destacam-se o Dublin Core, Data Documentation Initiative (DDI) e a International Organization for Standardization (ISO) 19115. Destaca-se aqui que 23% dos repositórios que possuem categorização por metadados não apresentam a definição exata de qual padrão utilizam, sendo categorizados como outros (Figura 8).

FIGURA 8: PADRÕES DE METADADOS DOS REPOSITÓRIOS



Fonte: Elaboração própria

O padrão Dublin Core está presente em 17% dos repositórios que apresentam informações sobre seus metadados. Segundo informações apresentadas no DCMI (2016), o padrão possui cinco grandes princípios: construção de forma aberta e sob consenso; participação e escopo internacional; neutralidade em seus propósitos e modelos de negócio; neutralidade na tecnologia e foco transdisciplinar. A organização e operação da iniciativa é formalizada através de uma série de regras que definem como os membros da iniciativa atuação no que diz respeito a suas responsabilidades e os processos de decisão sobre o padrão.

Já o padrão DDI é focado na descrição de dados provenientes das ciências sociais, comportamentais e econômicas (KRAMER; LEAHEY, 2012) e está presente em 14% dos repositórios que possuem esta característica. Este padrão é expresso em XML e contempla todo o ciclo de vida dos dados de pesquisa, desde a etapa de conceituação, passando pela análise e arquivamento dos dados. O projeto de criação do padrão DDI remonta ao ano de 1995, sendo que a primeira versão foi apresentada à comunidade em idos de 2000. Desde então o padrão passou por diversos processos de redesenho e a versão atual é a DDI Lifecycle 3.2, que possui características modulares e a capacidade de ser expandida (DATA DOCUMENTATION INITIATIVE, 2013).

Por fim, o padrão ISO 19115 é aplicado em 13% dos repositórios. O padrão foi atualizado pela última vez no ano de 2014 e seu foco é na descrição de informações geográficas. Segundo a ISO (2014), o padrão é aplicável a catalogação de todos os tipos de recursos, bem como serviços geográficos, conjunto de dados e características geográficas.

Ao se analisar os três padrões de metadados percebe-se que o Dublin Core possui características mais abrangentes em relação a sua utilização. O DDI também possui esta característica, mas o seu processo de criação foi focado em áreas específicas do conhecimento. Por fim, o padrão ISO 19115, embora possua a indicação de uso para qualquer tipo de documento, seu desenvolvimento foi focado em dados geográficos, o que indica que o padrão é o mais específico entre os três.

Repositórios de dados de pesquisa e a ciência aberta

A última categoria de análise avaliou o comportamento dos repositórios em relação aos princípios da ciência aberta declarados pelo Panton Principles (2011). O Panton Principles: Principles for Open Data in Science, foi elaborado por quatro pesquisadores, três do Reino Unido e um dos Estados Unidos, e é assinado por mais de 265 pesquisadores. O documento parte da ideia geral que a ciência é baseada na reutilização e discussão do conhecimento científico já publicado, e que este processo é mais efetivo se os dados das publicações estiverem abertos, ressaltando assim os princípios da ciência moderna. De acordo com Molloy (2011), a abertura dos dados significa a disponibilização livre na Internet com permissões para download, cópia, análise, reprocessamento e uso em software. Concluindo que os dados precisam ser disponibilizados em um local de domínio público. Após a definição de contexto e de conceitos, são propostos quatro princípios para a abertura dos dados de pesquisa (Quadro 1)

QUADRO 1: PANTON PRINCIPLES

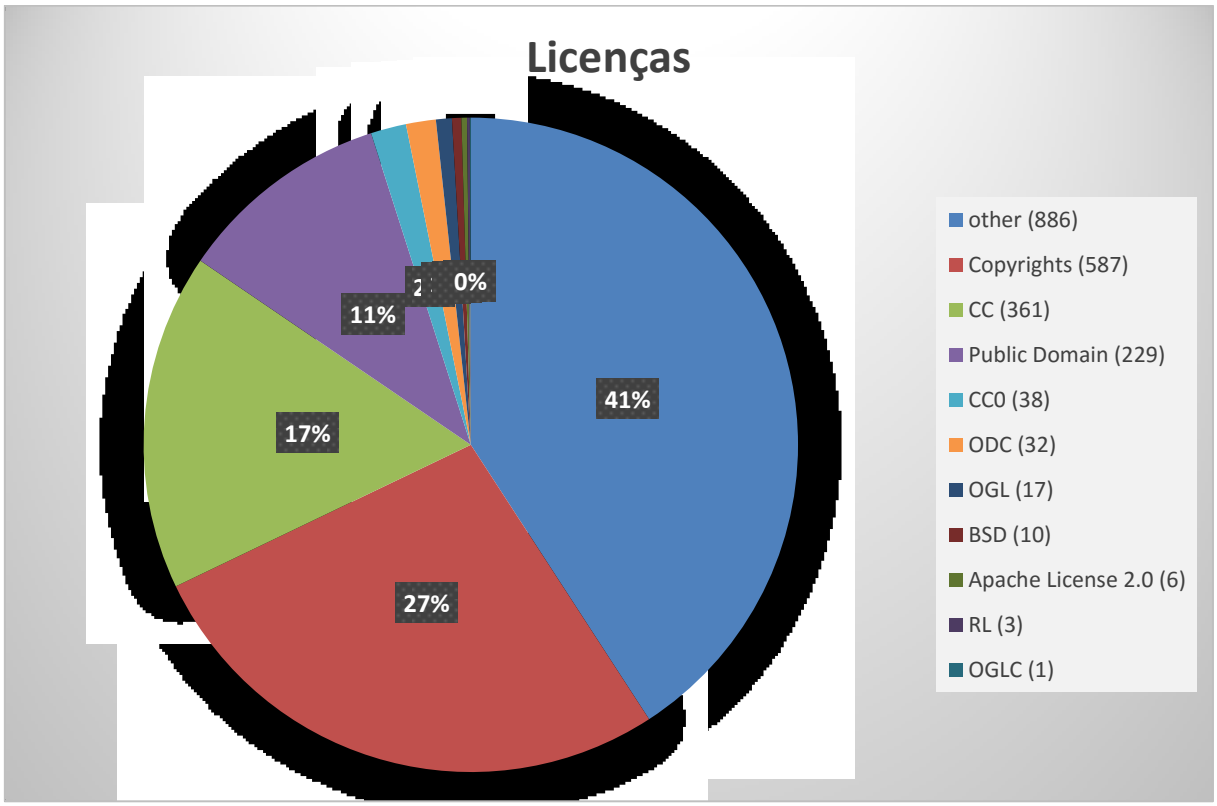
Princípios
1. É livre para uso e reuso.
2. É distribuído sem restrições
3. Utiliza a licença que determina que dados derivados compartilha a mesma licença de abertura.
4. É disponível em todo seu conjunto.
5. Quando possuir custo de reprodução, este deve ser razoável.
6. Preferencialmente deve ser apto para o acesso por meio da Internet e sem custos.
7. É disponível em formato conveniente e modificável.

Fonte: Elaboração própria

Nos princípios enunciados sobressai a perspectiva de abertura dos dados, no qual implica a disponibilização livre na Internet com permissões para *download*, cópia, análise, reprocessamento e uso em *software*. Para tanto, as permissões devem ser explicitamente declaradas por meio de uma licença adequada, conforme destacou Molloy (2011).

De forma geral, os repositórios analisados não foram satisfatórios em relação a apresentação de licença para a ciência aberta, pois 41% deles não apresentaram informações sobre a licença utilizada e 27% declararam utilizar a licença *copyright*. Apenas um quarto dos repositórios cumpriram a adequação proposta pelo Panton Principles, 17% utilizando licenças Creative Commons e 11% empregando licenças Public Domain (Figura 9).

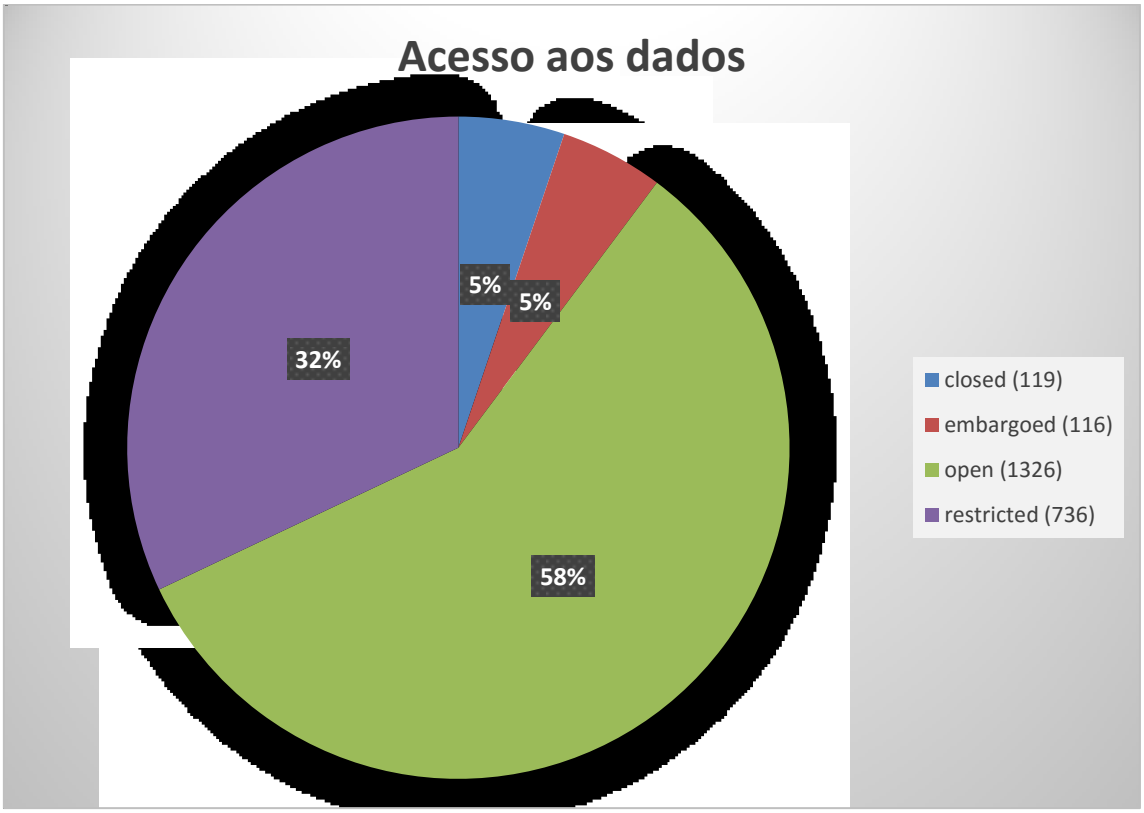
FIGURA 9: LICENÇAS UTILIZADAS NOS REPOSITÓRIOS DE DADOS DE PESQUISA



Fonte: Elaboração própria

Apesar das deficiências relacionadas às licenças de acesso e uso aos conteúdos, 58% dos repositórios analisados declararam permitir o acesso aberto direto e 5% somente após um período de embargo. Na contramão da ciência aberta, 37% dos repositórios declararam não promover o acesso aberto aos dados de pesquisa que armazenam (Figura 10).

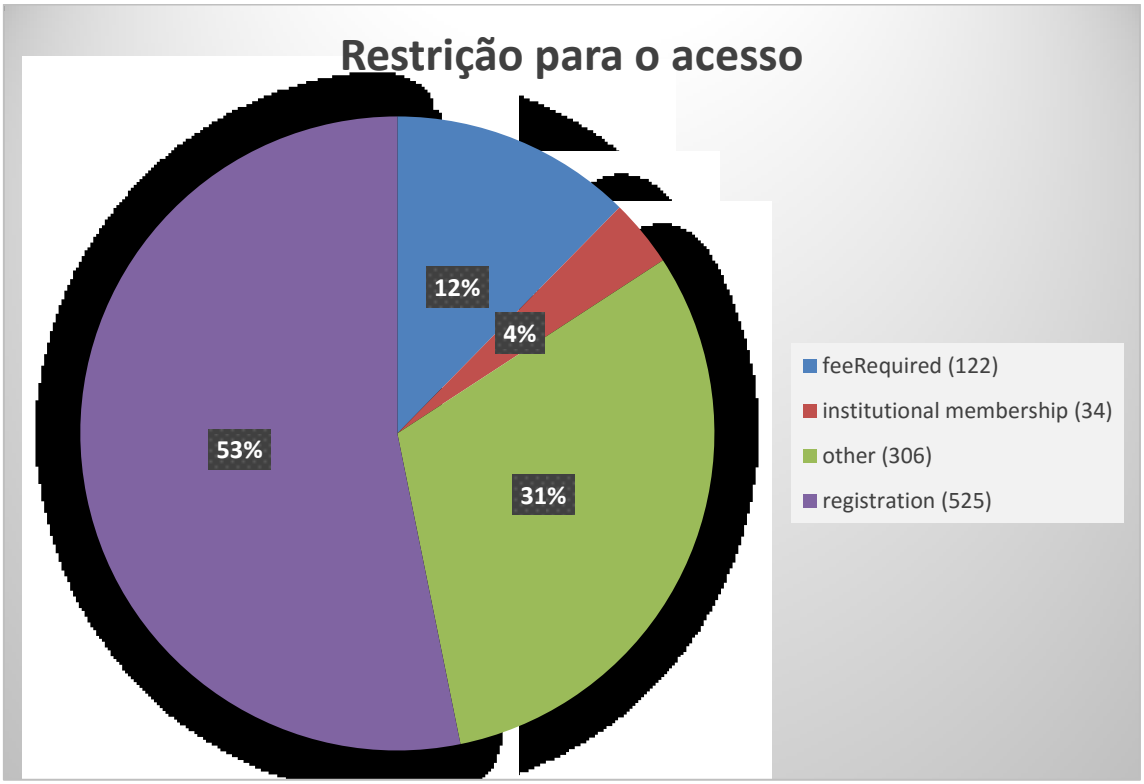
FIGURA 10: POLÍTICA DE ACESSO AOS DADOS ARMAZENADOS NOS REPOSITÓRIOS DE DADOS DE PESQUISA



Fonte: Elaboração própria

Parte das restrições impostas pelos repositórios relacionam-se ao registro dos usuários no sistema (53%) e à verificação da filiação institucional do usuário (4%). Além destas, observou-se que 12% das restrições implicam no pagamento de taxas para o acesso aos dados de pesquisa armazenados (Figura 11).

FIGURA 11: RESTRIÇÕES PARA O ACESSO AO CONTEÚDO DISPONÍVEL NOS REPOSITÓRIOS DE DADOS DE PESQUISA



Fonte: Elaboração própria

Considerações finais

Conclui-se, portanto, que o cenário mundial dos repositórios de dados de pesquisa aponta uma tendência para criação de repositórios temáticos, em vez de institucionais e aplicação de certificados específicos de acordo com as áreas do conhecimento, embora a certificação não seja necessariamente uma tendência. A ampla expressividade das Ciências da Saúde indica que suas práticas de financiamento, produção e comunicação de conhecimento científico são favoráveis ao contexto da ciência colaborativa e ao compartilhamento e reutilização de dados de pesquisa. No entanto, observou-se que metade dos repositórios analisados, independente da área do conhecimento, não licenciam seus registros de forma adequada para a ciência aberta, comprometendo sua ampla disseminação e reutilização. Por fim, a visão ampla acerca da disponibilidade de dados apresentada nesse artigo indica que para atingir os objetivos da ciência aberta há, ainda, que se obter avanços nessa área que é considerada um dos elementos essenciais para a democratização da ciência.

Referências bibliográficas

BOULTON, G. Reinveting Open Science for the 21st Centrury. In: Uma década de acesso

aberto na UMinho e no Mundo. Lisboa: Universidade do Minho, 2013. p. 239–250.

DATA DOCUMENTATION INITIATIVE. DDI Lifecycle 3.2. Disponível em: <http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/>. Acesso em: 6 out. 2016.

DCMI. Dublin Core Metadata Initiative (DCMI): About. Disponível em: <http://dublincore.org/>. Acesso em: 13 out. 2016.

GEZELTER, D. What, exactly, is Open Science?The OpenScience Project, 2009. Disponível em: <http://www.openscience.org/blog/?p=269>. Acesso em: 4 out. 2016

ISO. ISO 19115-1:2014 - Geographic information -- Metadata -- Part 1: Fundamentals. Disponível em: <http://www.iso.org/iso/catalogue_detail.htm?csnumber=53798>. Acesso em: 13 out. 2016.

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

MOLLOY, J. C. The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biol, v. 9, n. 12, p. e1001195, 6 dez. 2011.

OECD. Declaration on Access to Research Data from Public FundingOrganization for Economic Co-operation and Development (OECD), , 2004. Disponível em: <http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157#_ftn0>

Panton Principles. Disponível em: <http://pantonprinciples.org/>. Acesso em: 12 maio. 2016.

WALPORT, M.; BREST, P. Sharing research data to improve public health. The Lancet, v. 377, n. 9765, p. 537–539, 18 fev. 2011.