ARTIGO

Recebido em: 26/06/2017

Aceito em: 31/07/2018

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 23, n. 53, p. 52-63, set./dez., 2018. ISSN 1518-2924. DOI: 10.5007/1518-2924.2018v23n53p52

Repositórios de dados de pesquisa na Espanha: breve análise.

Research Data Repositories in Spain: short review

Fernanda Passini MORENO (fpassini@gmail.com) *

* Doutora em Ciência da Informação e Professora Adjunta (DE) pela Universidade de Brasília, com Pós-Doutorado na Universidad Carlos III de Madrid, Espanha.

Resumo

A pesquisa aqui relatada investigou as características dos dados de pesquisa no cenário espanhol. Foram selecionados repositórios exclusivamente espanhóis registrados no diretório re3data.org a fim de identificar o uso de sistemas de informação/infraestrutura, a tipologia de dados e metadados relacionados, bem como as áreas mais representativas na disponibilização dos dados. Trata-se de pesquisa descritiva e qualitativa. Após a definição da amostra, foram analisados aspectos relacionados à distribuição de áreas de conhecimento, ao *software* utilizado, ao padrão de metadados e aos tipos de conteúdo. A amostra selecionada apresenta elementos representativos para fomentar discussões com a literatura: maior ocorrência da área de Humanidades e Ciências Sociais; utilização do DSpace como *software* e Dublin Core como padrão de metadados. As discussões finais e sugestões de estudos futuros incluem, entre outros aspectos, verificação posterior dos tipos de conteúdo presentes de fato nos repositórios registrados no diretório e a relação entre área de conhecimento, tipo de conteúdo e padrão de metadados.

Palavras-chave: Ciência Aberta. Metadados. Registro de repositório de dados de pesquisa.

Abstract

This present research investigated the research data characteristics in the Spanish scenario. Exclusively Spanish repositories, registered at re3data.org were selected aiming at identifying the use of information/infrastructure systems, related data and metadata typology, as well as the most typical areas for data availability. It is a descriptive and qualitative research. After setting up the sample, aspects regarding the distribution of areas of knowledge, software used, metadata standards and kinds of content were analyzed. The selected sample shows important elements to foster discussions with the literature: a bigger Humanities and Social Science areas incidence; the use of DSpace as the software and Dublin Core as the metadata patter. Final discussions and recommendations for further studies include, among other aspects, subsequent verification of content type actually present in the directory registered and the relation between knowledge area, content type and metadata standards.

Key-words: Open Science. Metadata. Registry of Research Data Repositories.

v. 23, n. 53, 2018.
p. 52-63
ISSN 1518-2924

1 INTRODUÇÃO

A criação de infraestrutura e manutenção de repositórios de dados de pesquisa está em curso em diversos países e apresenta-se como um desafio tanto em termos de gestão quanto da representação dos dados ou conjuntos de dados que estão contidos nesses sistemas.

Diversas são as tecnologias disponíveis que permitem a colaboração à distância entre pesquisadores, além da geração, reunião e compartilhamento de uma grande quantidade de dados de pesquisa no âmbito da Ciência Aberta (*Open Science*).

Como observam Fecher e Friesike (2013), *Open Science* pode significar um conjunto de conceitos, "que vão desde o democrático direito de acesso ao conhecimento (por exemplo, acesso aberto às publicações), até a demanda por incluir o público na pesquisa (por exemplo, ciência cidadã) para o uso de ferramentas de colaboração e de compartilhamento."

Autores como Hernández-Pérez e Garcia-Moreno já alertavam, em 2013, para a necessidade de apresentação de um plano de gestão de dados de pesquisa exigidos por agências de fomento, como, por exemplo a *National Science Foundation* (NSF), a *Organisation for Economic Cooperation and Development* (OECD) e a Comissão Europeia, por meio do e-IRG *eInfrastructure Reflection Group.* O detalhamento das funções de plano de gestão de dados pode ser encontrado em Sayão e Sales (2016).

A NSF demanda que todos os projetos de pesquisa que solicitem financiamento devem apresentar um plano de gestão de dados de pesquisa com fim de compartilhamento. Em 2004, representantes governamentais de 34 países da OECD aprovaram uma "*Declaration on Access to Research Data from Public Funding*", reconhecendo que o acesso livre poderia maximizar o valor do investimento público na ciência e que restrições a esse acesso poderiam diminuir a qualidade e a eficiência da investigação e da inovação científica. Entre 2006 e 2007, a OCDE aprovou e publicou o documento "OECD *Principles and Guidelines for Access to Research Data from Public Funding*", que constitui um documento de referência nessa temática. Já o documento mais recente do grupo europeu *eInfrastructure Reflection Group* (e-IRG) destaca como boas práticas e recomendações para promover o acesso, a preservação e a curadoria de grandes quantidades de dados, entre outros aspectos:

> [...] garantir que os formatos de dados estarão padronizados e conterão informação suficiente sobre o dado (metadado), para permitir o uso global dentro de uma disciplina, transversalmente entre disciplinas, e para novas configurações de pesquisa, que possivelmente não estavam previstas no momento da criação dos dados [...] (NFRASTRUCTURE REFLECTION GROUP, 2015, p. 11)

Assim, o fim último da proposta de disponibilização dos dados de pesquisa – compartilhamento e reuso – passa necessariamente pela representação destes dados, em certa infraestrutura tecnológica compatível a nível mundial, juntamente com o aceite e colaboração de pesquisadores.

Olson, Zimmerman e Bos (2008) argumentam que uma das forças que leva os cientistas a colaborarem é que nem sempre uma única disciplina tem condições de responder às questões apresentadas e que as agências de fomento têm incentivado a colaboração de especialistas em diferentes áreas. Destacam que as novas tecnologias permitem, além de viabilizar a colaboração à distância, gerar, reunir e compartilhar uma grande quantidade de dados.

No contexto europeu, o projeto *Facilitate Open Science Training for European Research* (FOSTER), trabalha com a proposta de auxiliar os diferentes membros da comunidade científica (estudantes, pesquisadores, etc.) a promover a ciência de acesso aberto. Nesse sentido, a definição de *Open Science* é:

> [...] a prática da ciência de tal forma que outros possam colaborar e contribuir, sempre que os dados da pesquisa, notas de laboratório e outros processos de pesquisa estão disponíveis gratuitamente, em termos que permitam a reutilização, a redistribuição e reprodução da pesquisa e seus dados e métodos subjacentes. (FACILITATE OPEN SCIENCE TRAINING FOR EUROPEAN RESEARCH, 2016)

O termo “dados abertos”, apesar de relativamente recente, é usado em vários domínios como, por exemplo, no âmbito governamental. Para os propósitos deste estudo, se compreende que os dados de pesquisa abertos são aqueles que estão disponíveis gratuitamente na internet.

Para o grupo britânico que desenvolveu os *Principles for Open Data in Science* (PANTON) (MURRAY-RUST et al., 2010):

> Por dados abertos na ciência entendemos que são aqueles que estão livremente disponíveis na internet, permitindo a qualquer usuário fazer o *download*, copiar, analisar, re-processar, passá-los para um *software* ou usá-los para qualquer outra finalidade sem outras barreiras financeiras, legais ou técnicas além daquelas próprias para obter acesso à Internet.

Como observou Murray-Rust (2008), a Declaração de Budapeste[1], ao trazer a definição de “acesso livre”, já embutia os conceitos de que todo o conteúdo (o texto, os dados e os metadados) poderiam ser reutilizados para qualquer propósito sem necessidade de permissão explícita adicional. O autor é um dos membros do grupo que criou os *Panton Principles* e adotou a mesma estrutura da Declaração de Budapeste.[2]

Na pesquisa aqui relatada, foi realizada uma primeira aproximação para compreensão do cenário de dados de pesquisa no contexto espanhol, que já possui consórcios e grupos de estudos estabelecidos sobre o tema associados ao FOSTER, tais como o Maredata - *Red Española sobre Datos de Investigación en Abierto* (http://maredata.net/), Curatore - *Custodia y gestión digital de datos de investigación* (http://www.curatore.es/) e Datasea - *Datos Abiertos de Investigación / Open Research Data* (http://www.datasea.es/dt/).

A Espanha possui 84 universidades, oferta quase 4.000 cursos de mestrado e mais de 1.000 cursos de doutorado, mantendo uma média de 10.000 teses defendidas. Mesmo reconhecendo que o ciclo da pesquisa científica não é restrito às universidades, ilustra-se o cenário para dimensionar o potencial de criação, geração e uso de dados de pesquisa. Ainda como exemplo, no período de 2004 a 2014, o país duplicou a presença de publicações em revista indexadas na *Web of Science* (OBSERVATORIO IUNE, 2016).

Buscou-se, portanto, conhecer melhor o ambiente de compartilhamento dos dados de pesquisa espanhóis por meio de levantamento de repositórios de dados registrados no diretório *Registry of Research Data Repositories*[3], conhecido como *re3data*. O re3data abriga repositórios de dados de pesquisa de diversas áreas de conhecimento, apresentando-se como um registro global desse tipo de repositório. Segundo Zhi-Feng (2014), a Espanha possuía 10 repositórios registrados naquele diretório em 2014. Os dados aqui apresentados e discutidos foram coletados entre fevereiro e março de 2017, e atualmente são 20 os repositórios registrados como espanhóis. As questões de pesquisa que norteiam o trabalho são: os repositórios espanhóis registrados no re3data pertencem a quais áreas? Quais *software*s são utilizados? Quais tipos de conteúdo e padrões de metadados são utilizados?

A próxima seção apresenta os procedimentos metodológicos adotados. A seção 3 apresenta os dados discutidos à luz da literatura recente e o trabalho é finalizado com considerações preliminares que levantam algumas inquietudes e possibilidades de pesquisa a partir dos dados levantados.

2 METODOLOGIA

A abordagem metodológica desta pesquisa pode ser definida como qualitativa e descritiva. Segundo Patton (2007), o que faz uma pesquisa ser rotulada como qualitativa é o

---

[1] *Budapest Open Access Initiative*. Disponível em: http://www.budapestopenaccessinitiative.org/read Acesso em: 10 maio 2016.

[2] “*By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparables from gaining access to the internet itself*”. (BUDAPEST OPEN ACCESS INITIATIVE, 2002)

[3] *Registry of Research Data Repositories.* Disponível em: http://www.re3data.org. Acesso em: 28 jan. 2017.

esforço em produzir relevantes descrições detalhadas e interpretações particularizadas de pessoas e o meio social, linguístico, material, e outras práticas e eventos que os moldam e são moldados por eles.

De acordo com Bhattacherjee (2012), a pesquisa descritiva é utilizada quando se propõe a fazer observações cuidadosas e detalhadas de um fenômeno de interesse.

Estudos em comunidades específicas podem fornecer um olhar aprofundado sobre aquela realidade e métodos qualitativos podem prover *insights* sobre os problemas pesquisados, gerando novas ideias e questionamentos.

Através da estratégia de levantamento, buscou-se responder: os repositórios espanhóis registrados no re3data pertencem a quais áreas? Quais *software*s são utilizados? Quais tipos de conteúdo e padrões de metadados são utilizados?

Entre fevereiro e março de 2017 foi realizada uma consulta por países no diretório re3data (http://www.re3data.org) selecionando-se o país Espanha, constituindo-se assim o universo pesquisado. Para composição da amostra, foram selecionados apenas aqueles repositórios que fossem exclusivamente espanhóis e não participassem de consórcios internacionais (Quadro 1).

Os dados da amostra foram coletados utilizando-se os filtros da ferramenta de descoberta e a lista de resultados do diretório. Além disso, procedeu-se a consulta a cada repositório da amostra, a partir de uma lista de verificação (coleta estruturada) com as seguintes categorias: **infraestrutura**, com subcategorias: nome do repositório, endereço (tanto original quanto o localizador no re3data) e a descrição *software* e **conteúdo**, com subcategorias: área do conhecimento, padrão de metadados e tipos de conteúdo.

Quadro 1 – Universo e amostra

Repositório consórcio europeu/internacional	Repositório espanhol
1. *European Archive of Historical EArthquake Data - AHEAD* 2. *Consortium of European Social Science Data Archives – CESSDA* 3. *Constrained Local UniversE Simulations - CLUES* 4. *Gran Telescopio CANARIAS Public Archive - GTC Public Archive* 5. *International Service of Geomagnetic Indices - ISGI* 6. *Italian Center for Astronomical Archive - IA2* 7. *MultiDark Database - Multimessenger Approach for Dark Matter Detection* 8. *Phenol-Explorer - Database on Polyphenol Content in Foods* 9. *Spanish CLARIN K-Centre - Centro-K CLARIN* 10. *The European Genome-phenome Archive - EGA* 11. *World Values Survey – WVS* 12. *e-cienciaDatos*	1. *Analysis of the Interstellar Medium of Isolated Galaxies - AMIGA* 2. *CEACS Data Library - Biblioteca de Datos de CEACS* 3. *Centro de Investigaciones Sociologicas Data Bank y Estudios - CIS Data Bank* 4. *DIGITAL.CSIC* 5. *Dipòsit Digital de la Universitat de Barcelona Dades* 6. *Herschel Science Archive – HSA* 7. *MindBigData - The "MNIST" of Brain Digits* 8. *UPF Digital Repository - Recursos i dades primàries - e-Repositori upf.*

Fonte: Elaborado pela autora a partir dos dados de re3data (março de 2017)

A etapa de consulta a cada repositório da amostra buscou elementos que visavam: confirmar a validade do endereço; verificar se o repositório estava ativo, buscar reconhecer o *software* quando o valor do resultado era "*Unknown*" (desconhecido) e reconhecer o padrão de metadados quando não era apresentada a descrição no diretório re3data.

Na próxima seção os dados são discutidos à luz da literatura, consultada nas bases de dados disponíveis na Biblioteca da *Universidad Carlos III de Madrid* e *Google* Acadêmico.

3 ANÁLISE E DISCUSSÃO DOS DADOS

A seção apresenta os dados encontrados e a discussão divididas nas seguintes subseções: área do conhecimento, *software*, metadados e tipos de conteúdo.

3.1 Áreas de conhecimento

Os repositórios de dados espanhóis registrados no diretório re3data são oriundos de diferentes áreas do conhecimento e apresentam, consequentemente, heterogeneidade na infraestrutura que sustenta os distintos tipos de dados. As áreas de conhecimento sinalizadas no re3data são baseadas no esquema de classificação da *German REsearch Foundaton* (DFG), válido para o período 2016 a 2019.

O esquema é divido em quatro grandes áreas (disciplinas científicas): Humanidades e Ciências Sociais (HCS); Ciências da Vida (CV); Ciências Naturais (CN) e Ciências da Engenharia (CE) e apresenta ainda subdivisões em mais três níveis: área de pesquisa, comitê de revisão e assunto (DEUTSCHE FORSCHUNGSGEMEINSCHAFT, 2016)

Assim, por exemplo, para a disciplina "Humanidades e Ciências Sociais" (1), uma área de pesquisa seria "Humanidades" (11), "Filosofia" seria um dos comitês de (108), sendo subdivido nos seguintes assuntos: História da Filosofia (108-01); Filosofia Teórica (108-02) e Filosofia Prática (108-03).

A documentação do diretório re3data (RÜCKNAGEL et.al., 2015) informa que o campo "assunto" é obrigatório, mas não há restrição quanto ao número máximo de ocorrências, tampouco se a descrição deve se limitar às disciplinas científicas (grande área) ou especificar ao nível de assunto.

Como afirmam Gómez, Méndez y Hernández-Pérez (2016), os repositórios registrados no diretório são autodescritos e podem estar em até quatro grandes áreas do conhecimento, caso aplicável a repositórios multidisciplinares.

Em sua autodescrição os repositórios da amostra utilizam, na maioria, três dos quatro níveis disponíveis no esquema de classificação, isto é, o último nível "assunto" é registrado em apenas três casos. Um repositório duplicava a grande área e dois repositórios da amostra listavam subclasses inexistentes, provavelmente em função do sistema de classificação utilizado ter sido atualizado em 2016. O quadro 2 demonstra a distribuição por áreas encontradas na amostra.

Quadro 2 – Repositórios de dados espanhóis e áreas de conhecimento

ÁREA / REPOSITÓRIO	HCS	CN	CV	CE
Analysis of the Interstellar Medium of Isolated Galaxies - AMIGA		■		
CEACS Data Library - Biblioteca de Datos de CEACS	■			
Centro de Investigaciones Sociologicas Data Bank y Estudios - CIS Data Bank	■			
DIGITAL.CSIC	■	■	■	■
Dipòsit Digital de la Universitat de Barcelona Dades	■		■	
Herschel Science Archive – HSA		■		
MindBigData - The "MNIST" of Brain Digits			■	■
UPF Digital Repository - Recursos i dades primàries - e-Repositori upf	■	■	■	

Fonte: Dados da pesquisa.

Para contagem das ocorrências foi considerada apenas a grande área/disciplina científica. Nos repositórios da amostra há predominância da área de Humanidades e Ciências Sociais (HCS). Pode-se compreender como multidisciplinar o repositório do *Consejo Superior de Investigaciones Científicas,* DIGITAL.CSIC.

Dois repositórios de universidades abrangem mais de uma disciplina: *Dipòsit Digital de la Universitat de Barcelona Dades* e UPF *Digital Repository*, além do *MindBigData*, que possui características peculiares comentadas na próxima seção. Estudo recente que se baseia no re3data como fonte de dados (KINDLING et. al., 2017), revela que a distribuição de

grandes áreas é centrada em Ciências Naturais seguido de perto por Ciências da Vida, o que pode ser percebido ainda hoje na navegação por assunto da ferramenta de descoberta.

A ferramenta do diretório re3data, no entanto, não permite levantar as intersecções entre as áreas, o que pode vir a limitar a compreensão sobre o reuso de dados por distintas comunidades. Este fato, somado à autodescrição de assunto - que não limita área e subáreas de conhecimento - constitui um desafio ao constituir uma amostra para pesquisa, como relatam Gómez, Méndez y Hernández-Pérez (2016).

3.2 *Softwares*

A consulta às páginas dos repositórios, além da checagem da descrição, forneceu os seguintes dados: três repositórios usam o *DSpace*, um utiliza o *Dataverse* e quatro não apresentaram elementos suficientes para sua determinação.

Encontrar o uso do *DSpace* como *software* para repositório de dados é natural em função da experiência de institutos de pesquisa e universidades na utilização de um *software* consolidado: no diretório *Open Doar* estão registrados 1480 repositórios que usam esta plataforma. Lançada no início dos anos 2000 é um sistema de gerenciamento de produção científica digital e, segundo Rodrigues et. al. (2004)[4] "é um sistema em código aberto 1) com uma arquitetura de *software* simples mas eficaz, (...) 4) muito direccionado para o acesso aberto à publicação académica." Além disso, Amorim et.al. (2016) ao compararem múltiplos aspectos de plataformas de gerenciamento de dados de pesquisa, apresentam entre as vantagens do sistema a representação de metadados estruturados e a compatibilidade com o protocolo *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH).

O *Dataverse,* criado pelo Instituto de Ciências Sociais Quantitativas (IQSS) da Universidade de Harvard, informa na página do projeto que:

> [...] é uma aplicação web de código aberto para compartilhar, preservar, citar, explorar e analisar dados de pesquisa. (...) Um repositório do Dataverse é uma instalação do software, que hospeda vários arquivos virtuais chamados Dataverses. Cada dataverse contém conjuntos de dados e cada conjunto de dados contém metadados descritivos e arquivos de dados (incluindo documentação e código que acompanham os dados) (...) dataverses também podem conter outros dataverses." (THE DATAVERSE PROJECT, 2018, documento não paginado)

Kindling et. al. (2017), relatam que 75% dos repositórios registrados no diretório re3data em 2015 não informavam ou não nomeavam o *software* base: somados, *DSpace* e *DataVerse* representam 4,8% do total. Uma possível resposta é dada pelos próprios autores: "O pequeno número de tipos de *software* registrados re3data é provavelmente devido ao nosso processo de indexação manual. Em muitos casos o *software* do repositório não é explicitamente indicado no *site* e como consequência não indexado". Os autores encontraram maior uso do *DSpace*, o que é confirmado com dados de 2016 por Flathers, Kenyon e Gessler. Esses últimos advogam pelo uso do modelo *Open Archival Information System* (OAIS) para preservação e manutenção de acesso aos dados em longo prazo, requisito atendido pelo *DSpace* mas não pela maioria dos outros sistemas para repositórios (op.cit, 2016, p. 4).

O relatório do grupo espanhol "*Depósito y Gestión de datos en Acceso Abierto*" (GRUPO DE TRABAJO DE "DEPÓSITO Y GESTIÓN DE DATOS EN ACCESO ABIERTO" DEL PROYECTO RECOLECTA, 2012) relata que foram encontrados casos de sucesso de uso do *DSpace* em várias áreas, com custos adequados para implementação.

A navegação nos repositórios que utilizam *DSpace* mostrou que quando o repositório criava a comunidade "dados de pesquisa", o acesso aos dados era facilitado, caso dos repositórios da Universidade de Barcelona (*Dipòsit Digital de la Universitat de Barcelona Dades*) e da Universidade Pompeu Fabra (UPF). Essa última estrutura uma comunidade para dados de pesquisa, mas cria coleções por departamentos. Acrescenta, ainda, "tipo de documento" como filtro na ferramenta de navegação e pode-se buscar por *dataset,* ou conjunto de dados.

[4] Documento não paginado.

Já o repositório do DIGITAL.CSIC, possui indicadores robustos: 143.293 registros disponíveis. Porém, trata-se de repositório institucional que mescla uma variedade de recursos produzidos por pesquisadores de 150 institutos de pesquisa. A organização das coleções e comunidades inviabiliza o acesso aos conjuntos de dados.

Em relação aos outros quatro repositórios da amostra, estes apresentavam na descrição termos como "*other*", "*unknow*" ou sem preenchimento na indicação de *software.* Dentre eles, um potencialmente adota um sistema de gestão de conteúdo de código aberto *Open CMS* (*Content Management System*) - dedução em função da sigla presente na URL (*Uniform Resource Locator*).

O acesso aos dados de pesquisa, nesses casos, foi dificultado pela ausência e/ou limitação de ferramenta de busca, pelo engessamento da navegação estruturada ou pela total ausência de dados, como no caso do *MindBigData* (Figura 1), que de fato é apenas uma página web com animações e hiperligações para a rede social *Linkedin* e para empresas de *software*.

Figura 1 – Site do MindBigData

Fonte: http://www.mindbigdata.com/ (2017)

3.3 Padrões de metadados

Na ferramenta de descoberta do diretório, ao acessar a aba de filtros, são fornecidas as seguintes indicações para padrões de metadados: 2 ocorrências para *Data Documentation Initiative* (DDI)[5], 3 ocorrências para *Dublin Core* (DC) e uma ocorrência para *Flexible Image Transport System* (FITS)[6], relativas ao universo de 20 repositórios. Das seis ocorrências de indicação de metadados, 3 não pertenciam à amostra (DDI, FITS e 1 DC).

A consulta à descrição individual no diretório somada à consulta aos repositórios em si revelou 3 ocorrências na amostra para o *Dublin Core*, ou seja, havia um repositório que utilizava DC, mas não estava registrado como tal. As três ocorrências encontradas são relativas aos repositórios que usam DSpace, Um maior uso do DC confirma o encontrado por Gómez, Méndez y Hernández-Pérez (2016) ao analisarem os repositórios de Humanidades e Ciências Sociais registrados no re3data.

A partir de uma lista pré-definida, o re3data admite a seleção entre 33 tipos de padrões de metadados para descrever o repositório que será contido no registro, além da opção "outros". A lista pré-definida é baseada nos padrões referenciados no *Digital Curation*

[5] "A Iniciativa de Documentação de Dados (DDI) é um padrão internacional para descrever os dados produzidos em pesquisas e outros métodos de observação nas ciências sociais, comportamentais, econômicas e da saúde. O DDI é um padrão gratuito que pode documentar e gerenciar diferentes estágios no ciclo de vida dos dados de pesquisa, como conceituação, coleta, processamento, distribuição, descoberta e arquivamento. A documentação dos dados com o DDI facilita o entendimento, a interpretação e o uso - por pessoas, sistemas de software e redes de computadores". (DDI ALIANCE, 2018, documento não paginado)

[6] O Sistema Flexível de Transporte de Imagem (FITS) é um formato de arquivo de dados de imagem para codificação de dados astronômicos, usado para o transporte, análise e arquivamento de conjuntos de dados científicos. (FLEXIBLE IMAGE TRANSPORT SYSTEM, 2017)

Centre (DCC)[7], que indica 39 padrões de metadados para os mais diferentes domínios, ou "metadados disciplinares."

Para Willis, Greenberg, White (2012, p. 1505), que discutem os objetivos dos metadados para descrição dos dados de pesquisa:

> Apesar dos benefícios notáveis, a proliferação de esquemas de metadados específicos para disciplina também contribuiu para o estabelecimento de barreiras artificiais à descoberta de dados e reutilização por outras disciplinas. Essas barreiras, frequentemente associadas à semântica de metadados e estruturas de dados, interferem no progresso científico junto a linhas multidisciplinares, interdisciplinares e transdisciplinares. Juntas, as barreiras podem interferir no progresso que apoia nossa compreensão contemporânea da ciência.

O relatório do grupo espanhol "*Depósito y Gestión de datos en Acceso Abierto*" (GRUPO DE TRABAJO DE "DEPÓSITO Y GESTIÓN DE DATOS EN ACCESO ABIERTO" DEL PROYECTO RECOLECTA, 2012), expõe a importância dos metadados para dados de pesquisa para o cumprimento de funções relacionadas a "gestão e administração, a preservação, a descrição, a disseminação dos dados e à recuperação dos dados. " O relatório do grupo espanhol, além de indicar alguns padrões de metadados mais genéricos – como o *Metadata Encoding and Transmission Standard* (METS) e o próprio DC, lista os requisitos mínimos para identificação de um *dataset*, o que não encontrou eco na amostra pesquisada.

3.4 Tipos de conteúdo

Os 15 tipos de conteúdo aceitos para descrição no diretório pertencem a um vocabulário controlado oriundos do *PARSE.Insight - Permanent Access to the Records of Science in Europe*. Segundo Bhattacharya (2013), *PARSE.Insight* foi um projeto focado na análise de necessidades e mapeamento de infraestrutura necessária para apoiar a persistência, acessibilidade e entendimento/reusabilidade de dados e documentação científica, de acordo com as diretrizes do *Open Archival Information Systems* (OAIS). O projeto foi desenvolvido pela *European Alliance for Permanent Access* (APA). No Quadro 3 são descritos os tipos de conteúdo e os exemplos de tipos de arquivos definidos pelo projeto.

Quadro 3 – Tipos de conteúdo e exemplos de tipos de arquivos.

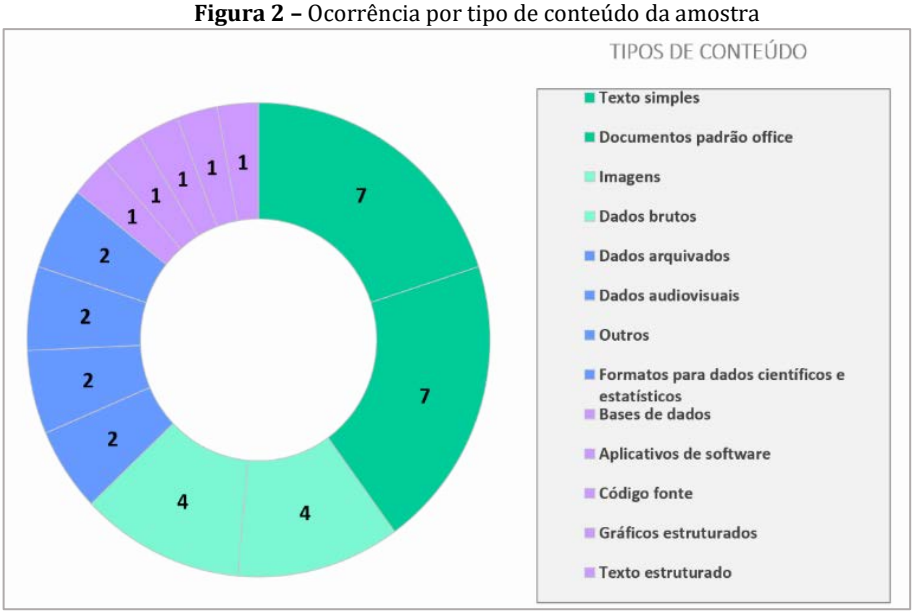
Tipos de conteúdo	**Exemplos de tipos de arquivos**
Documentos padrão *office*	Documentos de texto, planilhas, apresentações
Dados baseados em rede	Páginas da *web*, correio eletrônico, histórico de *chats*, etc.
Bases de dados	*DBASE, MS Access, Oracle, MySQL*, etc.
Imagens	*JPEG, JPEG2000, GIF, TIF, PNG, SVG*, etc.
Gráficos estruturados	*CAD, CAM, 3D, VRML*, etc.
Dados audiovisuais	*WAVE, MP3, MP4, Flash*, etc.
Formatos para dados científicos e estatísticos	*SPSS, FITS, GIS*, etc.
Dados brutos	Saída específica do dispositivo
Texto simples	*TXT* em várias codificações
Texto estruturado	*XML, SGML*, etc.
Dados arquivados	*ZIP, RAR, JAR*, etc.
Aplicativos de *software*	Ferramentas de modelagem, editores, Ambiente de Desenvolvimento Integrado (*Integrated Development Environment*), compiladores, etc.
Código fonte	Linguagem de *script, Java, C, C++, Fortran,* etc.
Dados de configuração	Parâmetros de configuração, *logs*, arquivos de biblioteca
OUTROS	-

Fonte: RÜCKNAGEL et. al. (2015), traduzido pela autora.

[7] O DCC possui uma série de recursos para curadoria - manutenção, gestão e preservação a longo prazo - de dados de pesquisa, disponíveis em: http://www.dcc.ac.uk/resources

A amostra revelou dois tipos principais de conteúdo: texto simples e documentos padrão *office*, com 7 ocorrências cada, seguidos de dados brutos e imagens, com quatro ocorrências cada, como se vê na Figura 2.

Figura 2 – Ocorrência por tipo de conteúdo da amostra



Fonte: Dados da pesquisa

Os dados encontrados são similares aos de Kindling et. al. (2017), ao explorarem a distribuição de tipos de conteúdo em todo o re3data.Os autores ao relataram mais ocorrências de texto simples, documentos padrão *office* e imagens para Humanidades e Ciências Sociais, áreas que tem mais presença na amostra estudada, dados que diferem de Gómez, Méndez y Hernández-Pérez (2016), que para as mesmas áreas encontraram maior incidência para os dados científicos e estatísticos. A ocorrência relativamente alta na amostra de imagem como tipo de dado pode ser explicada pelo processo de digitalização, comuns nessas áreas.

4 CONSIDERAÇÕES FINAIS

O escrutínio possível a partir de uma pequena amostra e por meio de uma análise qualitativa, revelou que em relação às áreas de conhecimento, a distribuição de disciplinas científicas não dista do conjunto de repositórios registrados. A estratégia de coletar e tratar as ocorrências de descrição do campo "assunto" mostrou-se acertada por proporcionar uma melhor visão sobre as áreas de conhecimento, reconhecer padrões e revelar inconsistências na descrição dos repositórios da amostra. Sobre esse último tópico, cabe observar ainda que:

- dados duplicados podem estar sendo utilizados ao adotar-se como método contagens de ocorrências sem o devido tratamento dos dados (por exemplo, duplicação de grande área ou subárea);
- análises futuras por subáreas do conhecimento deverão levar em conta a atualização do sistema de classificação utilizado no re3data. Este poderá se tornar um tipo comum de análise por pesquisadores do campo da Comunicação Científica, principalmente a partir do crescimento dos dados de pesquisa disponíveis. Deste modo, sugere-se considerar as possíveis reclassificações adotadas pelo sistema, nem sempre refletidas no diretório, além das particularidades da organização do conhecimento presentes no sistema alemão.
- a autodeclaração acerca da área de conhecimento/disciplina científica pode levar à vieses nos dados coletados diretamente no re3data. As coletas que não considerarem multidisciplinaridade ou se valerem apenas das ocorrências indicadas nos filtros da ferramenta de descoberta, poderão levar à classificação estanque em 4 áreas.

Em relação aos *softwares* utilizados, os repositórios espanhóis acompanham as tendências de uso (quando indicados) do *DSpace* e *Dataverse*, que somados, representam metade das ocorrências da amostra. As universidades presentes na amostra apresentaram os dados de forma mais organizada e acessível através da criação de uma comunidade específica para abrigá-los, o que pode ser uma tendência para unidades de informação que já utilizam e tenham a intenção de agregar dados de pesquisas em seus repositórios.

A ocorrência de "outros" e/ou "desconhecido" como indicação de *software*, escondeu a admissibilidade, pelo diretório, de páginas web sem dados de pesquisa disponíveis, ou de páginas estáticas que só eventualmente davam acesso ao texto completo e/ou que não forneciam ferramentas de busca. Uma possível resposta é a pressão dos mandatos de agências de fomento locais e/ou órgãos supranacionais. Mais recentemente, em maio de 2016, o Conselho de Ministros da União Europeia acordou que todas as publicações financiadas com dinheiro público sejam de acesso livre em 2020, incluindo as revistas científicas e dados de pesquisa. Essas políticas podem estar levando os institutos de pesquisa e universidades à "disposição" dos dados, mas não à promoção do acesso e recuperação dos mesmos.

No tópico relativo aos metadados e tipos de conteúdo, o predomínio do DC como padrão de metadados é diretamente ligado ao predomínio do *DSpace* como *software* base dos repositórios. Note-se, porém, que sem a consulta ao repositório em si, a amostra a partir dos dados do diretório revelaria mais uma inconsistência: dois usos de DC e não três. Ainda que haja recomendações de uma estrutura básica na descrição de dados de pesquisa na Espanha, ela não é seguida nem na descrição do diretório nem nos repositórios em si.

Os tipos de conteúdo encontrados revelam que documentos de texto ou padrão *office* tem maior presença na amostra, mas somente uma outra estratégia de pesquisa revelaria se os tipos de conteúdo indicados já estão contidos nos referidos repositórios ou são aqueles passíveis de serem admitidos, posto que a indicação do tipo não passa por checagem pela equipe do diretório.

Alguns questionamentos poderão ensejar pesquisas futuras: É possível estabelecer uma relação direta entre área de conhecimento e tipo de conteúdo que ela produz? Se admitirmos que a navegação no *DSpace* e *DataVerse* são mais amigáveis, a verificabilidade dessa relação teria como universo apenas 5% de todos os repositórios registrados no re3data.

Considerando a grande disponibilidade de *software*s e de padrões de metadados interoperáveis para diversos domínios, é possível que as comunidades ou áreas de pesquisa estejam encontrando problemas ao "se reconhecerem" em padrões de metadados? Uma estratégia seria investigar os atores do processo, pois como argumenta Borgman (2012), os curadores de dados, bibliotecários e outros envolvidos na gestão de dados podem oferecer uma coleção de dados do seu ponto de vista - mas que não seja percebida pelos destinatários como tal. Por outro lado, um investigador pode manter coleções de materiais e dados diversos, sem se dar conta de quão valiosos estes podem ser.

Os dados encontrados e as possibilidades de investigação anteriormente mencionadas poderão vir a contribuir para o panorama dos dados de pesquisa, que já encontra reflexos na Ciência da Informação brasileira, com seminários e números especiais de periódicos dedicados ao tema.

REFERÊNCIAS

AMORIM, R.C. et al. A comparison of research data management platforms: architecture, flexible metadata and interoperability. **Univ. Access. Inf. Soc.**, Heidelberg v. 16, n. 4, jan. 2017. Disponível em: <https://www.springerprofessional.de/en/a-comparison-of-research-data-management-platforms-architecture-/11691508>. Acesso em: 08 jun. 2017.

BHATTACHARYA, S. G. P. K. Insight into issue of Permanent Access to the Records of Science in Europe (PARSE.Insight). In: INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES (ICDL) 2013: VISION 2020, 1, 2013, New Delhi. **Conference papers.** New Delhi: Teri, 2013. p. 7 - 13.

BHATTACHERJEE, A. **Social science research:** principles, methods, and practices. Florida: USF, 2012.

BORGMAN, C. L. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology**, New York, v. 63, n. 6, p. 1059-1078, jun. 2012. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/full>. Acesso em: 15 jul. 2016.

BUDAPEST OPEN ACCESS INITIATIVE. 2002. Disponível em: <http://www.budapestopenaccessinitiative.org/read>. Acesso em: 10 de maio de 2016.

DATAVERSE. **About the project**. Disponível em: <http://dataverse.org/about>. Acesso em: 19 jul. 2018.

GERMAN RESEARCH FOUNDATION. **Classification of Scientific Disciplines, Research Areas, Review Boards and Subject Areas (2016-2019).** Bonn: DFG, 2016. Disponível em: <http://www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/amtsperiode_2016_2019/fachsystematik_2016-2019_en_grafik.pdf>. Acesso em: 03 mar. 2017.

DDI ALLIANCE. **Document, Discover and Interoperate.** 2018. Disponível em: <https://www.ddialliance.org>. Acesso em: 14 jul. 2018.

eINFRASTRUCTURE REFLECTION GROUP. **Best Practices for the use of e-Infrastructures by large-scale research infrastructures**. The Hague: e-IRG, 2015. Disponível em:< http://e-irg.eu>. Acesso em: 20 de jun de 2016.

FECHER, B.; FRIESIKE, S. **Open Science:** One Term, Five Schools of Thought. Berlim: German Data Forum (RatSWD), 2013. Disponível em: <www.ratswd.de/dl/RatSWD_WP_218.pdf >. Acesso em: 25 de jun. de 2015.

FLATHERS, E.; KENYON, J., GESSLER, P. E. A service-based framework for the OAIS model for earth science data management. **Earth Sci Inform**, v.10, n. 3, p. 383–393, set. 2017. Disponível em: <http://link.springer.com/article/10.1007/s12145-017-0297-3>.

FLEXIBLE IMAGE TRANSPORT SYSTEM. The Astronomical image and table format. 2017. Disponível em: <https://fits.gsfc.nasa.gov/fits_home.html>. Acesso em: 19 jul. 2018.

FACILITATE OPEN SCIENCE TRAINING FOR EUROPEAN RESEARCH. 2016. Disponível em: <https://www.fosteropenscience.eu/about> . Acesso em: 13 maio 2016.

GÓMEZ, N.; MÉNDEZ, E.; HERNÁNDEZ-PÉREZ, T. Social Sciences and Humanities research data and metadata: a perspective from thematic data repositories. **El profesional de la información**, v. 25, n. 4, p. 545-555, 2016. Disponível em: <http://dx.doi.org/10.3145/epi.2016.jul.04>. Acesso em: 30 jan. 2017.

GRUPO DE TRABAJO DE "DEPÓSITO Y GESTIÓN DE DATOS EN ACCESO ABIERTO" DEL PROYECTO RECOLECTA (Espanha). **La conservación y reutilización de los datos científicos en España:** informe del grupo de trabajo de buenas prácticas. Madrid: FECYT, 2012. Disponível em: <www.fecyt.es>. Acesso em: 20 fev. 2017.

HERNÁNDEZ-PEREZ, T.; GARCÍA-MORENO, M. A. Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios. **El profesional de la información**, v. 22, n.3, p. 259-263, maio/jun. 2013. Disponível em: <http://www.elprofesionaldelainformacion.com/contenidos/2013/mayo/10.html>. Acesso em: 15 dez. 2016.

KINDLING, M. et al. The landscape of research data repositories in 2015: A re3data Analysis. **D-Lib Magazine,** v. 23, n. 3/4, mar./abr. 2017. Disponível em: <http://www.dlib.org/dlib/march17/kindling/03kindling.html#21>. Acesso em: 10 mar 2017.

MURRAY-RUST, P. Open Data in Science. **Serials Review**, v. 34, n. 1, p. 52-64, jan. 2008. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/00987913.2008.10765152>. Acesso em: 20 fev. 2017

OBSERVATÓRIO IUNE. **Informe IUNE 2016:** Actividad investigadora de la universidad española. Madrid: INAECU, 2016. Disponível em: <www.informes.iune.es/Informe%20IUNE%202016.pdf>. Acesso em: 20 abr. 2017

OLSON, G.M.; ZIMMERMAN, A.; BOS, N. **Scientific collaboration on the Internet**. Cambridge: MIT Press, 2008.

MURRAY-RUST, Peter et al. Panton Principles: Principles for open data in science. 2010. Disponível em: <https://pantonprinciples.org/about/>. Acesso em: 02 maio. 2017.

PATTON, M. Q. Qualitative Evaluation. In: LEWIS-BECK, M. S.; BRYMAN, A.; LIAO, T. F. (Eds.) **Encyclopedia of Social Science Research Methods**. Thousand Oaks, CA: SAGE Publication, 2007. Disponível em: <http://sk.sagepub.com/reference/socialscience/n781.xml>. Acesso em: 10 mar. 2017.

RODRIGUES, E. et al. RepositóriUM: implementação do DSpace em português: lições para o futuro e linhas de investigação. In: CONFERÊNCIA DA ASSOCIAÇÃO PORTUGUESA DE SISTEMAS DE INFORMAÇÃO, 5, Lisboa, 2004. **Actas**. Lisboa: APSI, 2004. Disponível em: <http://repositorium.sdum.uminho.pt/handle/1822/679>. Acesso em: 26 fev. 2017.

RÜCKNAGEL, J. et. al. **Metadata Schema for the Description of Research Data Repositories**. Disponível em: <http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1397899:6/component/escidoc:1398549/re3data_schema_documentation_v3_0.pdf>. Acesso em: 13 mar. 2017.

SAYÃO, L. F.; SALES, L. F. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Inf. Inf.,** Londrina, v. 21, n. 2, p. 90-115, maio/ago. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939/20122>. Acesso em: 16 maio 2017.

THE DATAVERSE PROJECT. **About**. 2018. Disponível em: <https://dataverse.org/about>. Acesso em: 18 jul. 2018.

WILLIS, C.; GREENBERG, J.; WHITE, H. Analysis and synthesis of metadata goals

for scientific data. **Journal of the American Society for Information Science and Technology**, New York, v. 63, n. 8, p. 1505-1520, jun. 2012. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22683>. Acesso em: 08 de mar. De 2017.

ZHI-FENG, Z. The Study on Distribution of Research Data Repositories Based on Resources Directory Websites. In: INTERNATIONAL CONFERENCE ON ECONOMIC MANAGEMENT AND TRADE COOPERATION, 2014, Xian. **Proceedings**. Paris: Atlantis Press, 2014. p. 204 - 211. Disponível em: <http://www.atlantis-press.com/php/download_paper.php?id=11753>. Acesso em: 25 mar. 2017.

Editores do artigo: Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.