XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-8 – Informação e Tecnologia

MAPEAMENTO DOS METADADOS PARA DADOS CIENTÍFICOS

Ana Carolina Simionato (Universidade Federal de São Carlos - UFSCar)

MAPPING OF METADATA FOR SCIENTIFIC DATA

Modalidade da Apresentação: Comunicação Oral

Resumo: A coleta de dados sempre foi recorrente no âmbito científico, no entanto, novas políticas e solicitações pelas principais agências de fomento à pesquisa estão sendo feitas aos pesquisadores e às instituições de pesquisa quanto as questões de gestão, preservação e acesso aos dados gerados pelas pesquisas. A preocupação para o armazenamento e representação desses dados, configura-se dentre os estudos da área de Ciência da Informação. Nesse contexto, torna-se vital o questionamento de quais são os metadados utilizados na representação os dados científicos? A partir disso, foi feito o recorte temático para esse trabalho, delimitando-se a identificar os metadados que estão sendo utilizados na descrição de dados em repositórios de dados científicos e que utilizam como plataforma o *DSpace*. A pesquisa é de natureza teórico-aplicado e qualitativa, e em relação ao método do trabalho, essa pesquisa é classificada como exploratória. Como resultado é apresentado o mapeamento dos metadados de 50 repositórios de dados científicos, identificados no *Registry of Research Data Repositories* (re3data). Conclui-se pelo mapeamento e pela análise realizada que a predominância para representação dos *datasets* é o uso dos metadados: *dc.title*, *dc.subject*, *dc.contributor.author*, *dc.description*, *dc.date.accessioned*, *dc.date.available*, *dc.date.issued*, *dc.type*, *dc.rights*, *dc.rights.uri*, e *dc.language.iso*, sendo que a arquitetura para descrição de dados não está sendo aperfeiçoada para especificidade de cada temática dos repositórios, e também, destaca-se a importância dos metadados no ciclo de vida dos dados científicos para a disponibilização e reuso dos dados.

Palavras-Chave: Descrição de datasets; Metadados; Repositórios de dados; Dublin Core; Dados científicos.

Abstract: Data collection has always been applicant scientific, however, new policies and requests by the major funding agencies for research are being made to researchers and research institutions about the issues of management, preservation and access to data

generated by the research. The concern for storage and representation of this data, is one of the studies in the field of information science. In this context, it becomes vital to the question of what are the metadata used in scientific data? From that, it was made the cut for this work, bordering to identify the metadata that is being used in the description of data in scientific data repositories using DSpace platform. The research is theoretical and applied nature and qualitative, and in relation to the method of work, this research is classified as exploratory. As a result, the mapping of metadata for scientific data repositories 50, identified in the Registry of Research Data Repositories (re3data). We conclude by mapping and analysis held that the prevalence for representation of datasets is the use of metadata: dc title, dc. subject, dc. contributor. author, dc. description, dc. accessioned, dc. available, dc. issued, dc. type, dc. rights, dc.rights.uri, and dc.language.iso, and the architecture for data description is not being optimized for specific character of each of the thematic repositories, and highlights the importance of metadata in science data lifecycle for the provision and reuse of data.

Keywords: Description of datasets; Metadata; Scientific data repositories; Dublin Core; Scientific data.

1 INTRODUÇÃO

Produzidos por inúmeras formas, seja por humanos ou por máquinas, os dados, em sua grande parte, nem sempre estão estruturados de forma coerente para serem disponíveis aos usuários, como o que acontece com o fenômeno *Big Data*, conjunto aos fatores de velocidade, volume, variedade e complexidade dos dados.

Os dados possuem uma estrutura mínima e se caracterizam por multidimensionar as possibilidades de acesso, uso e reuso, tornando assim, um crescente campo de estudo à Ciência da Informação. Essas questões também podem ser relacionadas a representação, manutenção e organização da informação tratadas pelas disciplinas de catalogação, classificação e indexação. Assim, a busca por um tratamento informacional hábil e por uma padronização, continua sendo a mesma de tempos remotos. (BACA, 2016).

Nesse viés, o contexto acadêmico também se depara com a dificuldade de tratamento para os dados, principalmente pela atribuição de valores em que os dados possuem e na rápida necessidade de estarem disponíveis. O valor é adquirido pela eficácia que os dados possuem na prevenção de fraudes de pesquisa, como principalmente, para multidimensionar as possibilidades de uso e reuso dos dados, discorrendo à sustentabilidade dos dados.

As definições para dados de pesquisa podem variar de acordo com a comunidade de pesquisa. Para as áreas de biológicas e saúde, são definidos como os materiais registrados e comumente aceitos na comunidade científica, conforme a documentação e pelos resultados de pesquisa. (NATIONAL INSTITUTES OF HEALTH, 2012). Para National Science Foundation (2015) os dados são determinados pela comunidade e pelo gerenciamento de programas, incluindo os dados derivados de publicações, amostras, coleções físicas, *software* e modelos. A National Endowment for the Humanities (2017) define como os materiais gerados ou coletados durante a realização de pesquisas, fornecendo uma variedade de exemplares.

As publicações de dados científicos vêm sendo cobrados pelas maiores instituições de financiamento internacional, como a *National Science Foundation*, *National Institutes of Health*, *National Endowment for the Humanities*, *Economic and Social Research Council*, *Wellcome Trust* e mais recentemente, a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

No entanto, para que os dados científicos estejam disponíveis, é necessário um planejamento e gerenciamento eficiente, iniciado desde a confecção do *Data Management*

Plan (DMP) pelo pesquisador até o depósito dos dados durante toda a sua execução até a sua finalização, considerando os fatores éticos, segurança e confiabilidade dos dados digitais. Nesse contexto, a FORCE11 (2014) apresenta facetas para orientação do gerenciamento de dados, cujas facetas são fundamentadas pelos princípios FAIR, assim, os dados devem ser encontrados (*findable* - F) sendo identificáveis e persistentes, acessíveis (*accessible* - A) na medida de serem obtidos pode máquinas e humanos, interoperáveis (*interoperable* - I) sendo executáveis por máquinas, utilizarem vocabulários compartilhados e/ou ontologias, e serem sinteticamente analisados e semanticamente acessíveis à maquina, e reusáveis (*re-usable* – R) que condizem a estarem suficientemente descritos e integrados com outras fontes de dados, possibilitando a citação adequada.

Consequentemente, o gerenciamento dos dados tornou-se uma fase essencial para o desenvolvimento de pesquisas, e ainda, para garantir que cada pesquisa esteja alinhada aos seus dados de origem. Nesse contexto, novos procedimentos estão sendo desenvolvidos e criados, como a preparação do armazenamento dos dados, como a implantação de repositórios de dados em universidades e áreas temáticas. Nesse contexto, Amorim et al. (2016) apresentou um panorama das plataformas de gerenciamento de dados que podem ser implementadas por uma instituição para apoiar o fluxo de trabalho de gerenciamento de dados, tais plataformas foram *DSpace*, *CKAN*, *Figshare*, *Zenodo*, *ePrints*, e *EUDAT*.

Ao iniciar a implantação do repositório de dados, "[...] as exigências sobre o nível de descrição e de atribuição de metadados devem ser identificadas desde o começo do seu projeto e revistas ao longo do ciclo de vida dos seus dados, sendo essa a essência de uma boa curadoria de dados" (SAYÃO; SALES, 2017, p. 28). Nesse sentido, os metadados destacam-se pela importância no ciclo de vida dos dados científicos, viabilizando a sua disponibilização e reuso dos dados, como já apontados anteriormente com os princípios FAIR.

Por essa razão, a representação de dados pode tornar-se uma grande e crescente preocupação dentre os estudos da área de Ciência da Informação, e da mesma forma, temas comuns da Biblioteconomia podem revitalizar-se em proporções maiores quando, o recurso a ser tratado é maximizado para um conjunto de dados.

Os metadados são conceituados na Ciência da Informação, como os valores e atributos referentes aos recursos informacionais. Nesse ponto, incita que o conceito de recurso informacional, também pode ser atribuído ao *dataset*, baseado na definição de Glushko (2014, p. 08, tradução nossa) "O recurso possui um sentido comum de 'qualquer coisa de valor que

pode apoiar a atividade orientada a um objetivo.'". Com a discussão de representação para *datasets*, torna-se vital o questionamento de quais são os metadados utilizados na representação os dados científicos?

Para tanto, foi feito o recorte temático para esse trabalho, delimitando-se a identificar os metadados que estão sendo utilizados na descrição de dados em repositórios de dados científicos e que utilizam como plataforma o *DSpace*, objetivando analisar uma arquitetura mínima de representação para dados em repositórios de dados científicos.

Justifica-se a delimitação desse trabalho para o uso do *Dspace* em consequência da amplitude temática que a definição dos metadados em repositórios de dados ocasiona. Nesse sentido, pressupõe-se que resultados apontaram para a utilização do *Dublin Core*, visto que plataforma e própria comunidade *Dspace* já o indicam para a utilização em sua instalação. No entanto, ressalta-se que o uso do *Dublin Core* não é uma obrigatoriedade, sendo possível fazer a configuração de outro padrão e a personalização dos metadados do próprio padrão, optando pelo seu formato simples ou qualificado, ou mesmo, sendo possível realizar adaptações com a inserção de outros metadados específicos, como exemplo na composição de perfis de aplicação.

2 PROCEDIMENTOS METODOLÓGICOS

Este estudo é de natureza teórico-aplicado e qualitativo, e em relação ao método de trabalho, essa pesquisa é classificada como exploratória. Aos procedimentos técnicos recorre-se a pesquisa bibliográfica, com o levantamento realizado em nível nacional e internacional em fontes de pesquisa primárias, secundárias e terciárias. Sendo o período selecionado transcorrente de 2000 a 2017.

As fontes utilizadas no levantamento bibliográfico foram: Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI), Portal de Periódicos da Capes, Google Acadêmico, *Scientific Electronic Library Online* (SciELO) e a *Web of Science*, tendo como resultados majoritários artigos, teses/dissertações, capítulos de livros e livros sobre a temática.

Além da pesquisa bibliográfica, foi realizado um mapeamento dos metadados baseados na proposta de Santos, Simionato e Arakaki (2014) de definição de metadados para recursos informacionais, denominada como metodologia BEAM. Para análise, foram

considerados os repositórios cadastrados no *Registry of Research Data Repositories* (re3data)[1], plataforma de âmbito global, que mantem cadastrado repositórios de dados científicos para auxiliar pesquisadores e instituições na indicação de envio de dados para o armazenamento, preservação e acesso aos dados de científicos.

Para localização dos repositórios que seriam analisados, foi criado um filtro apenas com repositórios que utilizam o *DSpace* e que resultou em 50 repositórios. Os repositórios encontrados são: *Khazar University Institutional Repository*, PRISM: *University of Calgary Institutional Repository*, *Scholar's Bank*, KU *ScholarWorks*, *Repositorio Institucional USIL*, *DSpace@MIT*, *Edinburgh DataShare Dipòsit Digital de la Universitat de Barcelona Dades*, DRYAD, USU *Institutional Repository*, *Woods Hole Open Access Server*, *Unisa Institutional Repository*, UPF *Digital Repository - Recursos i dades primaries*, RIT *Digital Media Library Repository*, *Digibug:Repositorio Institucional de la Universidad de Granada, Wittliff Collections*, *Texas State University*, *Digital Collections Repository*, LINDAT/CLARIN *repository*, MINDS@UW, *Research Data Online*, CLARINO *Bergen Center repository*, *ShareGeo open*, EDINA, GO-GEO, *DepositOnce*, *WormBase*, SIOR, *Data Repository for the University of Minnesota*, *ScholarsArchive*@OSU, *Open Research Exeter*, JEDI, DataSTORRE, IDEALS, DIGITAL.CSIC, *Apollo*, *Språkbanken*, *eCommons - Cornell's digital repository, Banco de Información para la Investigación Aplicada en Ciencias Sociales, Repositorio Institucional*, UA *Campus Repository*, CLARIN INL *Portal*, *Banff International Research Station for Mathematical Innovation and Discovery*, SPECTRa *Project*, CLARIN-PL, DataDOI, UPSpace, *Earth-prints Repository*, CLARIN.SI repository, *Datorium*, ILC-CNR *for* CLARIN-IT *repository*, *WorldWideMolecularMatrix* e *South African Data Archive*.

Assim, para localização dos metadados foi utilizado o filtro *type* com valor *dataset* e escolhido um conjunto de registros e extraídos os metadados utilizados. Após a identificação dos metadados, os resultados foram tabulados em uma planilha *Excel* com os metadados localizados e mapeados com a proposta do *crosswalk* de St Pierre e La Plant (1999). Para facilitar na visualização dos resultados, os dados foram mapeados a partir dos elementos do *Dublin Core*.

[1] Consulta disponível pelo link: http://www.re3data.org/search?query=&software%5B%5D=Dspace. Acesso em: 10 ago. 2017.

3 DESENVOLVIMENTO TEÓRICO

As dificuldades para a representação dos dados derivam da complexidade de suas interpretações, das várias possibilidades de ligação, e também, a relação de sustentabilidade que pode ser construída a partir da utilização e reutilização dos dados. Os dados são definidos como "[...] conjunto mínimo de símbolos que pode ser tomado como uma unidade de conteúdo, precisa ser identificado com o contexto a que pertence." (SANTOS; SANT'ANA, 2013, p. 202). Santos e Sant'ana (2013, p. 205) ainda continuam a pontuar que

> [...] uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que, mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação.

Ressalta-se que o tratamento e a preocupação com os dados dessas instituições refletem desde as formas de criação a sua divulgação, tendo como denominação, manutenção de dados (GLUSHKO, 2014). Da mesma forma, como já apontado anteriormente, o planejamento para o uso dos metadados deve ser o início da arquitetura do repositório de dados científicos. Definindo metadados, Pomerantz faz a comparação com um mapa e sua área territorial, os metadados "[...] são um meio pelo qual a complexidade de um objeto é representada de uma maneira mais simples." (POMERANTZ, 2015, p. 11, tradução nossa).

Os metadados possuem papel na infraestrutura para descrição de recursos informacionais. Segundo Buckland (2007, p. 3, tradução nossa) "Os metadados possuem dois componentes: um formato e um conjunto de valores." No entanto, Qin e Li (2013), consideram que a infraestrutura de metadados possuem três componentes: semântica, técnica e política. Sendo que a semântica está relacionada aos vocabulários, esquemas e modelos; os componentes técnicos operam e entregam os serviços de metadados; e as políticas estão relacionadas as diretrizes de melhores práticas, regras e políticas de governança das práticas dos metadados.

Qin, Ball e Greenberg (2012) evidenciam as características dos metadados para descoberta, controle, qualidade e uso de dados. No contexto de dados científicos, Qin et al. (2010, p. 128, tradução nossa) explicam que "O termo metadados científicos é frequentemente usado para se referir aos dados que descrevem os conjuntos de dados

coletados ou gerados a partir de pesquisas científicas". Os metadados são categorizados, segundo a *National Information Standards Organization* (RILEY, 2017) em quatro grupos:

- Metadados descritivos: usados para encontrar ou entender um recurso, além de promover a interoperabilidade. Constituem-se, por exemplo, de título, autor, assunto etc.;
- Metadados estruturais: mostram as relações entre as partes de um recurso, promovendo a navegação. Tratam, por exemplo, de sequência, posição na hierarquia etc.;
- Linguagens de marcação: integram metadados e marcações para outras características semânticas ou estruturais dentro do conteúdo, visando a navegação e a interoperabilidade. Entre suas propriedades, incluem-se lista, parágrafo, cabeçalho etc;
- Metadados administrativos: estes são divididos em três subcategorias:
    - Metadados técnicos: usados para codificar e compilar os arquivos, contribuindo com a interoperabilidade, a preservação e com o gerenciamento de objetos digitais. Eles tratam do tipo do arquivo, tamanho do arquivo, data de criação etc.;
    - Metadados de preservação: usados para gerenciar arquivos em longo prazo e, assim como os metadados técnicos, são essenciais para a interoperabilidade, a preservação e o gerenciamento de objetos digitais. Eles informam eventos de preservação e armazenamento (um dígito que representa a soma de dígitos corretos em dados transmitidos ou armazenados, usado para detectar erros nos dados) entre outros;
    - Metadados de direitos autorais: usados para informar os direitos autorais atrelados ao conteúdo, contribuindo com a interoperabilidade e o gerenciamento de objetos digitais. Eles tratam de termos de licença, entidade detentora dos direitos etc.

As categorias dos metadados auxiliam na estruturação dos sistemas informacionais, Tennant (2004) destaca como principais características para os metadados devem possuir, que são: versatilidade, extensibilidade, abertura e transparência, limite mínimo e máximo, gestão cooperativa, modularidade, hierarquia, granularidade e qualidade de seus valores.

Para isso, estruturas padronizadas para a representação do conteúdo informacional devem ser salientadas, tendo a finalidade da recuperação e acesso (ALVES; SANTOS, 2013). Sendo os padrões de metadados definidos por "[...] um conjunto estruturado, padronizado, codificado e predeterminado de elementos de metadados que serão utilizados na representação descritiva dos recursos informacionais, aplicações e ou compartilhamento de dados entre sistemas." (ALVES; SANTOS, 2013, p. 13). Entende-se, então, que o conjunto de metadados que será utilizado para descrever um recurso, bem como a ordem deles no registro da unidade de informação, é dependente do padrão adotado, pois este possui um determinado esquema de metadados.

As discussões sobre a descrição de dados científicos foram abordadas por White et al. (2008), Qin et al. (2010, p. 128), Qin, Ball e Greenberg (2012), Qin e Li (2013), Chen, Lin e Chen (2013) e Alam (2014).

Qin e Li (2013, p. 30) destacaram os principais elementos entre diversos padrões utilizados para descrição de dados científicos, em que se constatou que os elementos mais comuns foram os descritivos, genéricos e temporal. Na análise realizada por Qin e Li (2013, p. 30) os elementos que mais apareciam no contexto dos dados científicos foram: descrição e título com dez ocorrências, publicador, citação e cidade com oito ocorrências cada uma, referência, endereço e palavras-chave com sete ocorrências, e-mail, data, resumo, identificador, propósito, cidade, criador e código postal com seis ocorrências cada uma. Os metadados de comentários, edição, fonte, número de telefone, versão, estado ou província e *status* tiveram a ocorrência de cinco vezes cada nome e lugar de publicação tiveram quatro ocorrências em cada elemento.

Chen, Lin e Chen (2013) propuseram uma relação de metadados para descrição de dados científicos como: título (*dc:title*), título alternativo (*dcterms:alternative*), o responsável primário por criar o documento como criador (*dc:creator*), já para outras contribuições para criação de um recurso foi indicado pelos autores o uso do *dc:contributor*. O publicador foi atribuído para o metadado *dc:publisher*. A descrição do *dataset* e do item foram indicados para utilizar o metadado *dc:description*. O metadado *dc:type* foi indicado para representar o tipo. O metadado *dc:format* foi utilizado para representar o formato e o tamanho foi utilizado o *dcterms:extent*. O assunto foi utilizado o *dc:subject* e a cobertura espacial e temporal foram utilizados *dcterms:spatial* e *dcterms:temporal*, respectivamente. A data foi utilizado o *dc:date* para designar a data de submissão, de aceite e *updated* e avaliação. O metadado referente ao

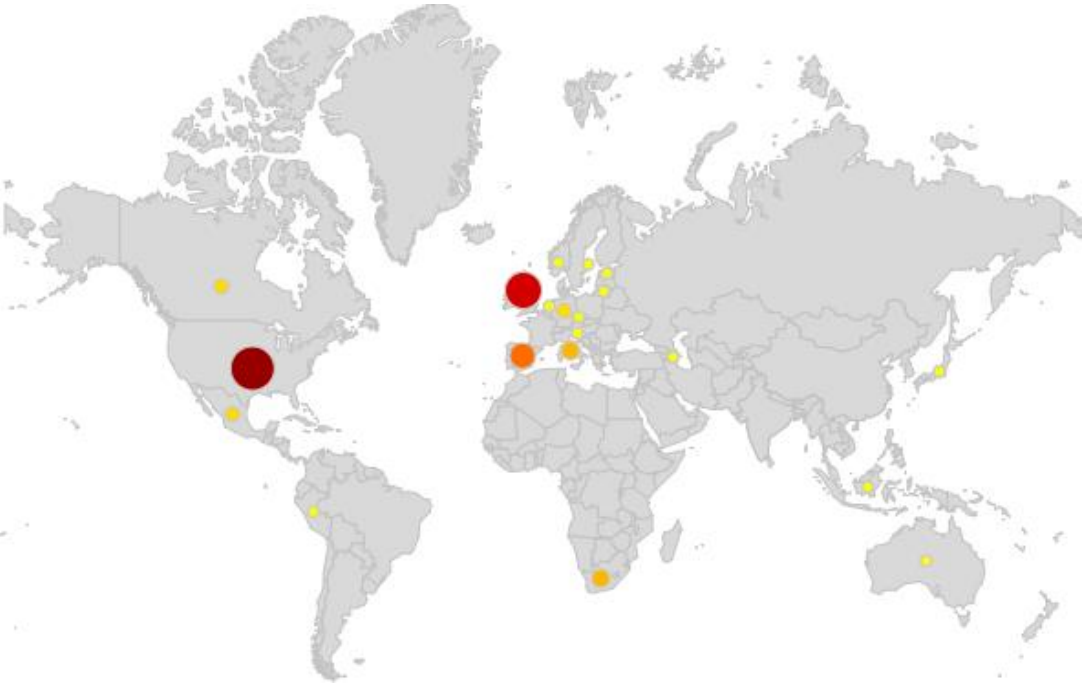
idioma foi utilizado *dc:language*, para designar a fonte foi utilizado o *dc:source* e para os relacionamentos foi utilizado o *dc:relation*. Os metadados sobre direitos, permissão de acesso aos metadados e ao download do item foram todos inseridos no *dc:rights.* O identificador do registro foi proposto utilizar o metadados *dc.identifier.uri*. Os metadados *dc:format*, *dcterms:extent* e *dc:date* são gerados automaticamente e os metadados *dc:title*, *dc:creator*, *dc:subject*, *dc:date* (*available*), *dc:language*. Os metadados sobre direitos, permissão de acesso, download do item, não foram considerados obrigatórios. Entretanto, as autoras destacaram que os tipos de metadados podem variar de acordo com o contexto e o tipo de dados que estão trabalhando, como no caso das áreas de meteorologia, geografia e geologia que necessitam de informações detalhados referentes à localização e o período em que os dados foram coletados. (CHEN; LIN; CHEN, 2013).

Alam (2014) apresentou a extensão do *Dublin Core* aplicada no repositório *Datorium*, gerenciado pela GESIS – *Leibniz Institute for the Social Sciences* para que possa descrever dados científicos. Segundo a autora o *Dublin Core* pode ser considerado como uma infraestrutura fundamental para publicações científicas e está evoluindo como uma infraestrutura comum para descrição de dados científicos (ALAM, 2014).

4 REPRESENTAÇÃO DE DADOS CIENTÍFICOS

Dentro do objetivo elucidado por esse trabalho, os 50 repositórios que utilizam o *Dspace* para descrição e armazenamento de *datasets* científicos, apresentam uma grande concentração na Europa e nos Estados Unidos, como ilustrado pela figura 1.

Figura 1: Distribuição geográfica de repositórios de dados científicos que utilizam *Dspace*



Fonte: Elaborado pela autora.

A figura 1 é ilustrada pela quantidade de repositórios de dados científicos que utilizam *Dspace*, do vermelho escuro ao amarelo, representam por tonalidade maior consecutivamente a maior proporção de repositórios disponíveis. No caso, os repositórios de dados em grande parte estão alocados nos Estados Unidos (EUA) com 15 repositórios, ou seja, 23,8%. O Reino Unido apareceu em segundo lugar com 17,45% dos repositórios (11 repositórios).

A Espanha apareceu em terceiro lugar com cinco repositórios, ou seja, 7,93% do total. Em seguida, a África do Sul e Itália tiveram 4,76%, ou seja, três repositórios cada um. Com 3,17% a Alemanha, o Canadá e o México foram identificados dois repositórios cadastrados no contexto desta pesquisa. Países como Austrália, Azerbaijão, Estônia, Indonésia, Japão, Lituânia, Noruega, Países Baixos, Peru, República Tcheca, Eslovênia e Suécia foram representados por um repositório cada um, ou seja, representaram ao todo 19,04%. Alguns repositórios estavam relacionados a contextos que extrapolavam um país, tendo como escopo a União Europeia, por exemplo, que contou com sete repositório que representaram 11,11% e de âmbito internacional, que foi representado por um repositório, ou seja, 1,58%.

Cada repositório teve extraído os dados de nomeação, país, orientação para algum esquema de descrição, como exemplo, o esquema da *Digital Curation Centre* que é baseado

no padrão de metadados *Dublin Core*, ou mesmo, o próprio padrão *Dublin Core*. Do registro de cada *dataset* completo foi extraído todos os metadados.

Os resultados e ocorrências foram sumarizados no quadro 1, em metadados mapeados e comparados com os quinze elementos do padrão *Dublin Core*.

Quadro 1: Número de ocorrências e metadados do mapeamento dos repositórios de dados científicos

DC Element	**Metadados mapeados (número de ocorrências)**
Title	dc.title (50), dc.title.alternative (1), dc.title.subtitle (1), dc.title.project (1)
Creator	dc.creator (5)
Subject	dc.subject (27), dc.subject.classification (3), dc.subject.other (2), dc.subject.ddc (1), dc.subject.lcsh (1)
Description	dc.description.refereed (1), dc.description (35), dc.description.sponsorship (5), dc.description.abstract (12), dc.description.tableofcontents (2), dc.description.department (1), dc.description.provenance (2), dc.description.collectioninformation (1)
Contributor	dc.contributor.author (29), dc.contributor.editor (1), dc.contributor.authorall
Publisher	dc.publisher (13), dc.publisher.corporate (1), dc.publisher.collection (1), dc.publisher.faculty (1), dc.publisher.department (1), dc.publisher.institution (1)
Date	dc.date.accessioned (40), dc.date.available (41), dc.date.issued (40), dc.date.updated (1), dc.date.created (5), dc.date.completion (1), dc.data (11), dc.date.publicationyear (2)
Type	dc.type (31), dc.type.embargo (1), dc.type.version (1)
Format	dc.format (2), dc.format.extent
Identifier	dc.identifier (3), dc.identifier.uri (46), dc.identifier.citation (5), dc.identifier.issn (1), dc.identifier.doi (3), dc.identifier.other (6)
Source	dc.source (1), dc.source.uri (1)
Language	dc.language.iso (22)
Relation	dc.relation (2), dc.relation.ispartof (5), dc.relation.uri (1), dc.relation.ispartofseries (2), dc.relation.requires (1),
Coverage	dc.coverage.spatial (6), dc.coverage.temporal (2), dc.coverage.spatialcounties (1), dc.coverage.spatialstates (1)
Rights	dc.rights (32), dc.rights.uri (28), dc.rights.accessRights (8), dc.rights.holder (1)
-	dc.extension (1), dc.notes (2)

Fonte: Elaborado pela autora.

Ao todo, foram identificados 69 metadados utilizados, variando entre o *Dublin Core* simples e qualificado. O metadado *dc.title* (título) foi considerado em todos *datasets* analisados. Foram identificados outras variações para o título como *dc.title*;

dc.title.alternative (título alternativo); *dc.title.subtitle* (subtítulo); *dc.title.project* (título do projeto).

Em relação à atribuição de autoria o metadado *dc.contributor.author* (autor) ficou em nono lugar aparecendo em 29 *datasets*. Em alguns *datasets* a atribuição de autoria variou em *dc.contributor.editor* (editor); *dc.contributor.authorall* (autoria de todos os participantes); *dc.contributor* (contribuidor); *dc.contributor.other* (outro contribuidor); *dc.contributor.affiliation*; *dc.contributor.department* e *dc.creator*, entretanto não tiveram muita representatividade. Outro ponto a se considerar foi que em alguns *datasets* utilizaram a designação de autoria em outros metadados como no título ou em assunto.

O metadado *dc.subject* (assunto) apareceu em 27 *datasets* e teve como variações *dc.subject.classification* (Classificação); *dc.subject.other*; *dc.subject.ddc*; *dc.subject.lcsh*. Nota-se que alguns repositórios optaram por utilizar no controle de seus vocabulários, estruturas hierárquicas como a Classificação Decimal de Dewey (CDD) e o *Library of Congress Subject Headings* (LCSH).

O metadado *dc.description* (descrição) apareceu em 35 *datasets*, em seguida o metadado *dc.description.abstract* apareceu em 12 *datasets*. Outras atribuições para designar algum tipo de descrição foram: *dc.description.refereed*; *dc.description*; *dc.description.sponsorship*; *dc.description.tableofcontents*; *dc.description.department*; *dc.description.provenance*; e *dc.description.collectioninformation*.

Em relação à designação do publicador, o metadado que mais apareceu foi o *dc.publisher* (publicador) utilizado em 13 *datasets*. As variações do metadado publicador foram utilizados: *dc.publisher.corporate*; *dc.publisher.collection*; *dc.publisher.faculty*; *dc.publisher.department*; e *dc.publisher.institution*.

Os metadados relacionados à data (*dc.date*), tiveram grande representatividade aparecendo em 41 *datasets* o metadado *dc.date.available* (data de disponibilização) e em 40 *datasets* os metadados *dc.date.accessioned* (data de aceitação) e o *dc.date.issued* (data de publicação). Outras variações identificadas fora: *dc.date.updated*; *dc.date.created*; *dc.date.completion*; *dc.date.publicationyear*. Destaca-se que em alguns *datasets* utilizaram o metadado *dc.date* (data) sem qualificador para designar a data de um recurso informacional. Nesse ponto, destaca-se que a falta de especificidade do metadado poderá acarretar em problemas de interoperabilidade com outros sistemas e até mesmo, dificultar o usuário ao

recuperar e entender se a data referida é uma data de publicação, data de depósito, data de revisão, entre outras possíveis datas.

O metadado *dc.type* (tipo) apareceu em 31 *datasets* e teve algumas variações como, *dc.type.embargo* e *dc.type.version*. Em todos os casos que esse metadado foi descrito, o termo do metadado foi *dataset*, referido ao tipo de recurso informacional.

Os metadados relacionados à identificadores como *dc.identifier*; *dc.identifier.uri*; *dc.identifier.citation*; *dc.identifier.issn*; *dc.identifier.doi*; e *dc.identifier.other*, foram utilizados em alguns *datasets*, mas o que teve maior representatividade foi o metadado *dc.identifier.uri* (identificador URI) com 46.

O metadado idioma (*dc.language.iso*) apareceu em 22 *datasets*. Já os metadados relacionados aos direitos (autorais e de acesso) tiveram maior representatividade em *dc.rights*, aparecendo em 32 *datasets* e o metadado, *dc.rights.uri* aparecendo em 28 *datasets*. Foram ainda utilizados em menor quantidade os *metadadosdc.rights.accessRights* e *dc.rights.holder*.

Também foram identificados metadados relacionados ao formato como *dc.format*; *dc.format.extent*; *dc.format.mimetype*, fonte (*dc.source* e *dc.source.uri*), os metadados de relação a outros itens, *dc.relation*; *dc.relation.ispartof*; *dc.relation.uri*; *dc.relation.ispartofseries*; *dc.relation.requires*; além dos metadados de cobertura (*dc.coverage.spatial*; *dc.coverage.temporal*; *dc.coverage.spatialcounties*; e *dc.coverage.spatialstates*), não tiveram representatividade em relação aos outros metadados. Algumas instituições criaram os metadados *dc.extension* (1 *dataset*) e *dc.notes* (2 *datasets*).

5 CONSIDERAÇÕES FINAIS

A coleta de dados no meio da pesquisa científica sempre foi recorrente, no entanto, novas políticas e solicitações pelas principais agências de fomento à pesquisa estão sendo feitas quanto as questões de gestão, preservação e acesso aos dados científicos.

Dessa forma, o gerenciamento de dados científicos permite ao pesquisador descobrir, interpretar, utilizar e reutilizar seus dados como de outros pesquisadores. A sustentabilidade dos dados, agrega valor ao momento inicial da pesquisa até as vias de sua publicação, pois infere diretamente na avaliação e verificação pelos pares, possibilitando que eles possam verificar e ainda, que obtenham resultados publicados, evitando possíveis fraudes.

Ao mesmo tempo, a rotina de pesquisa científica torna-se outra, a partir do momento que novos procedimentos são inclusos com a criação do *Data Management Plan*, que o proponente deverá desde o início da pesquisa refletir na preservação de seus dados, incluindo processos incomuns como o armazenamento, organização e representação de *datasets*.

Destaca-se assim, os metadados, como os dados estruturados que descrevem, explicam, identificam, localizam e representam um recurso. Isto é, os metadados permitem que os dados possam ser encontrados e interpretados, por meio de identificadores únicos e persistentes. Para tanto, os padrões de metadados auxiliam no controle de atributos e valores em sistemas informacionais, como os repositórios de dados.

No caso, foi verificado que o padrão de metadados *Dublin Core* ainda possui grande influência para a descrição de recursos digitais, tornando-se um dos principais padrões para a representação de dados científicos. A partir do mapeamento e da análise realizada, a predominância para representação dos *datasets* em repositórios de dados que utilizam o *Dspace* é o uso dos metadados: *dc.title*, *dc.subject*, *dc.contributor.author*, *dc.description*, *dc.date.accessioned*, *dc.date.available*, *dc.date.issued*, *dc.type*, *dc.rights*, *dc.rights.uri*, e *dc.language.iso*, sendo que a arquitetura para descrição de dados não está sendo aperfeiçoada para especificidade de cada temática dos repositórios, e também, destaca-se a importância dos metadados no ciclo de vida dos dados científicos para a disponibilização e reuso dos dados.

Destaca-se que mesmo o *Dspace* possibilitando a configuração de outro padrão de metadados, o *Dublin Core* está presente nos 50 repositórios de dados analisados. Ainda que pela complexidade que os dados apresentam, o *Dublin Core* simples atende as principais demandas de descrição, e que em alguns casos, a partir da análise dos repositórios de dados foi constatado a ocorrência de algumas especificações e qualificações do nível descritivo do Dublin Core.

Além disso, a partir da análise da literatura realizada, foi ratificado que os metadados mais utilizados nos repositórios de dados, são os mesmos sugeridos pelos autores: Qin e Li (2013) e Chen, Lin e Chen (2013).

Foi constatado que os repositórios do re3data que utilizam o *DSpace*, plataforma *open source*, apenas o Peru, na América do Sul, está presente. Concentrando-se a maior quantidade nos Estados Unidos e no Reino Unido, com diversas iniciativas na União Européia.

Considera-se que a representação de dados, ainda é um tema emergente e necessita de estudos concentrados em sua real aplicação e aproximação para a diversidade de áreas acadêmicas e científicas. Nesse sentido, aponta-se como trabalhos futuros, identificar a área temática dos repositórios analisados e realizar mapeamentos com padrões específicos às suas áreas do Conhecimento.

REFERÊNCIAS

ALAM, A. W. Dublin Core Metadata for Research Data: lessons learned in a real-world scenario with Datorium. **International Conference on Dublin Core and Metadata Applications**, p. 10, 2014.

ALVES, R. C. V.; SANTOS, P. L. V. A. DA C. **Metadados no domínio bibliográfico**. Rio de Janeiro: Intertexto, 2013.

AMORIM, R. C. et al. A comparison of research data management platforms: architecture, flexible metadata and interoperability. **Universal Access in the Information Society**, 11 jun. 2016.

BACA, M. (ED.). **Introduction to metadata**. 3. ed. Los Angeles: Getty Research Institute, 2016.

BUCKLAND, M. K. Description and search: metadata as infrastructure. **Brazilian Journal of Information Science: Research Trends**, v. 0, n. 0, p. 3–14, 2007.

CHEN, H.; LIN, Y.; CHEN, C. Approaches to building metadata for data curation. **International Conference on Dublin Core and Metadata Applications**, p. 4, 2013.

FORCE11. **Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0**. Disponível em: <https://www.force11.org/fairprinciples>. Acesso em: 13 ago. 2017.

GLUSHKO, R. J. **The discipline of organizing: core concepts edition**. Sebastopol, EUA: O'Reilly Media, 2014.

NATIONAL ENDOWMENT FOR THE HUMANITIES. **Data Management Plans for NEH Office of Digital Humanities**. Disponível em: <https://www.neh.gov/files/grants/>. Acesso em: 13 ago. 2017.

NATIONAL INSTITUTES OF HEALTH. **NIH Data Sharing Policy and Implementation Guidance**. Disponível em: <https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm>. Acesso em: 13 ago. 2017.

NATIONAL SCIENCE FOUNDATION. **Planning data for researchers: information on writing a data management paragraph for NSF**. Disponível em: <http://www.ru.nl/research-

information-services/data-management/planning-research/funder-requirements/nsf/>. Acesso em: 13 ago. 2017.

PIERRE, M. S.; LAPLANT, W. P. Issues in crosswalking content metadata standards. **Information standards quarterly**, v. 11, n. 1, p. 01-16, 1999.

POMERANTZ, J. **Metadata**. Cambridge, Massachusetts ; London, England: The MIT Press, 2015.

QIN, J.; BALL, A.; GREENBERG, J. Functional and architectural requirements for metadata: supporting discovery and management of scientific data. **International Conference on Dublin Core and Metadata Applications**, p. 10, 2012.

QIN, J.; LI, K. How portable are the metadata standards for scientific data? A proposal for a metadata infrastructure. **International Conference on Dublin Core and Metadata Applications**, p. 10, 2013.

RILEY, J. **Understanding Metadata: what is metadata, and what is it for?**National Information Standards Organization (NISO), , 2017. Disponível em: <http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf>

SANTOS, P. L. V. A. DA C.; SANTANA, R. C. G. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, v. 42, n. 2, p. 199–209, 2013.

SAYÃO, L. F.; SALES, L. F. Curadoria digital e dados de pesquisa. **AtoZ: novas práticas em informação e conhecimento**, v. 5, n. 2, p. 67, 9 jan. 2017.

TENNANT, R. A bibliographic metadata infrastructure for the twenty-first century. **Library Hi Tech**, v. 22, n. 2, p. 175–181, jun. 2004.

WHITE, H. C. et al. The Dryad Data Repository: a Singapore framework metadata architecture in a DSpace environment. **International Conference on Dublin Core and Metadata Applications**, p. 6, 2008.