Ricardo César Gonçalves Sant'Ana
Moisés Lima Dutra
Guilherme Ataíde Dias
Organizadores

ANAIS
WIDaT 2018

Organização do WIDaT 2018

- **Organização Geral:**

  Guilherme Ataíde Dias (PPGCI-UFPB) - Coordenador geral do evento

  Moisés Lima Dutra (PPGCIN-UFSC) - Vice-coordenador

- **Coordenador da Comissão Científica:**

  Ricardo César Gonçalves Sant'Ana (PPGCI-UNESP)

- **Comissão científica**

  Adilson Luiz Pinto (PPGCIN-UFSC)

  Ana Alice Baptista (Universidade do Minho, Portugal)

  Ana Carolina Simionato (PPGCI-UFSCar)

  Angela Maria Grossi de Carvalho (PPGCI-UNESP)

  Bernardina Maria Juvenal Freire de Oliveira (PPGCI-UFPB)

  Cristian Berrío-Zapata (PPGCI-UFPA)

  Dalton Lopes Martins (FCI-UnB)

  Denysson Axel Ribeiro Mota (PPGB-UFCA)

  Douglas Dyllon Jeronimo de Macedo (PPGCIN-UFSC)

  Ed Porto Bezerra (PPGI-UFPB)

  Edgar Bisset Alvarez (PPGCIN-UFSC)

  Edna Gusmão de Goés Brennand (MPGOA-UFPB)

  Edna Gomes Pinheiro (DCI-UFPB)

  Elaine Parra Affonso (FATEC-SP)

  Elvis Fusco (UNIVEM-Marília)

  Enrique Muriel Torrado (PPGCIN-UFSC)

  Evandro de Barros Costa (IC-UFAL)

  Fábio Paraguaçu (IC-UFAL)

  Fernando de Assis Rodrigues (PPGCI-UNESP)

  Gustavo Medeiros de Araújo (PPGCIN-UFSC)

  Henry Pôncio Cruz de Oliveira (PPGCI-UFPB)

  Joana Coeli Ribeiro Garcia (PPGCI-UFPB)

  José Eduardo Santarém Segundo (USP-FFCLRP)

  Leonardo Castro Botega (UNIVEM-Marília)

  Luana Farias Sales Marques (PPGCI-IBICT-UFRJ)

  Marckson Roberto Ferreira de Sousa (PPGCI-UFPB)

  Luís Fernando Sayão (CNEN)

  Marcelo Morandini (EACH-USP)

  Márcio Matias (PPGCIN-UFSC)

  Marcos Mucheroni (CBD-USP)

  Marynice de Medeiros Matos Autran (PPGCI-UFPB)

Maurício Barcellos Almeida (PPGGOC-UFMG)
Moisés Lima Dutra (PPGCIN-UFSC)
Plácida Leopoldina V. da Costa Santos (PPGCI-UNESP)
Pedro Luiz Pizzigatti Corrêa (POLI-USP)
Renata Baracho (PPGGOC-UFMG)
Ricardo César Gonçalves Sant'Ana (PPGCI-UNESP)
Robson Rodrigues Lemos (UFSC-Araranguá)
Rogério Ramalho (PPGCI-UFSCar)
Ryan Ribeiro de Azevedo (UFRPE-UAG)
Sandra de Albuquerque Siebra (PPGCI-UFPE)
Sandro Rautenberg (DECOMP-UNICENTRO)
Silvana Aparecida Borsetti G. Vidotti (PPGCI-UNESP)
Virginia Bentes Pinto (PPGCI-UFC)
Wagner Junqueira de Araújo (PPGCI-UFPB)
Zaira Regina Zafalon (PPGCI-UFSCar)

- **Coordenador do Cerimonial:**
  André Luiz Dias de França (PPGCI-UFPB)

- **Coordenador da Equipe Técnica Local:**
  Laerte Pereira da Silva Júnior (CCHLA-UFPB)

- **Equipe Técnica Local:**
  Adriana Alves Rodrigues (PPGCI-UFPB)
  Antonio Felipe dos Santos (MPGOA-UFPB)
  Débora Gomes de Araújo (PPGCI-UFPB)
  Pedro Augusto de Lima Barroso (PPGCI-UFPB)
  Pollianna Marys de Souza e Silva (PPGCI-UFPB)
  Renata Lemos dos Anjos (PPGCI-UFPB)

PADRÕES DE METADADOS PARA DESCRIÇÃO DE DADOS:
panorama dos repositórios de dados na América Latina

METADATA STANDARDS FOR DATA DESCRIPTION:
overview of Data Repositories in Latin America

Felipe Augusto Arakaki[1], Ana Carolina Simionato[2], Paula Regina Ventura Amorim Gonçalez[3], Plácida Leopoldina Ventura Amorim da Costa Santos[4]
(1) Instituto Federal de São Paulo (IFSP), Câmpus Itapetininga, Av. João Olímpio de Oliveira, 1561 - Vila Asem, Itapetininga - SP, 18202-000, felipe.arakaki@ifsp.edu.br.
(2) Universidade Federal de São Carlos (UFSCar), Rodovia Washington Luis, km 235 - São Carlos - SP, 13565-905, Departamento de Ciência da Informação, acsimionato@ufscar.br;
(3) Universidade Federal do Espírito Santo (UFES), Av. Fernando Ferrari, 514 - Goiabeiras, Vitória - ES, 29075-910, Departamento de Biblioteconomia, paula.goncalez@ufes.br,
(4) Universidade Estadual Paulista (UNESP), Av. Hygino Muzzi Filho, 737, Mirante, Marília, SP, 17.525-900, Programa de Pós-Graduação em Ciência da Informação, placida@marilia.unesp.br.

Resumo:
Os repositórios de dados de pesquisa são ambientes que oportunizam aos dados, produtos de pesquisa, serem acessados, compartilhados, utilizados e reusados, visto que, tais ambientes possibilitam a organização, o armazenamento e o acesso aos dados em diferentes formatos. A preocupação com a representação e recuperação dos conjuntos de dados é recorrente entre profissionais que tem a informação como objeto de trabalho e pesquisa. Assim, questiona-se quais são os padrões de metadados utilizados na representação dos dados científicos disponibilizados nos repositórios de dados na América Latina? Nesse contexto, o trabalho possui como objetivo analisar a formalização dos padrões de metadados utilizados para a descrição do conjunto de dados de pesquisa no âmbito dos repositórios de dados da América Latina registrados no *Registry of Research Data Repositories* (*re3data*). A pesquisa é de natureza teórico aplicada, com uma abordagem qualitativa no que se refere a representação da informação nos repositórios digitais de dados, para tanto optou-se pela pesquisa exploratória. Como resultado foi feito o mapeamento dos metadados dos 29 repositórios de dados científicos identificados no *re3data*. Ressalta-se que o reconhecimento para dados científicos e seu compartilhamento tem potencializado a criação de novas plataformas, *softwares* de gerenciamento e diferentes possibilidades para a descrição dos conjuntos de dados.
Palavras-chave: Descrição de conjuntos de dados. Metadados. Repositório de dados.

Abstract:
Research data repositories are environments that provide data, research products, are accessed, shared, used and reused, since such environments enable the organization, storage and access to data in different formats. The concern with the representation and retrieval of datasets is recurrent among professionals who have information as object of work and research. Thus, it is questioned what metadata patterns are used in the representation of the scientific data available in the data repositories in Latin America? In this context, the work aims to analyze the formalization of the metadata standards used to describe the set of research data within the Latin American data repositories registered in the Registry of Research Data Repositories (re3data). The research is of an applied theoretical nature, with a qualitative approach regarding the representation of the information in the digital repositories of data, for which we have opted for the exploratory research. As a result, the mapping of the metadata of the 29 research data repositories identified in the survey was done. It should be emphasized that the recognition of scientific data and their sharing has made possible the creation of new platforms, management software and different possibilities for the description of datasets.
Keywords: Dataset description. Metadata. Data repository.

1 INTRODUÇÃO

A preocupação para o efetivo gerenciamento de dados oriundos de pesquisas está sendo amplamente discutido pelas principais agências de fomento à pesquisa, visando o progresso científico além da contemporização de uma Ciência Aberta à todos os campos do conhecimento.

Esse movimento, torna-se salutar na concepção dos Repositórios de Dados. Os Repositórios de Dados de Pesquisa, se configuram como ambientes no auxílio a pesquisadores no que diz respeito à gerência, à disponibilização e ao acesso aos dados científicos, ações fundamentais, quando este ambiente permite e assegura o compartilhamento, o acesso e a reutilização dos dados, ações que reduzem sobremaneira o tempo e os gastos com nova coleta, como lhes chancelar validade diante a reprodução e replicação da pesquisa.

Vale destacar que os dados de pesquisa são definidos como:

> "[...] dados digitais são uma parte (descritiva) ou o resultado de um processo de pesquisa. Este processo abrange todas as etapas da pesquisa, que vão desde a geração de dados de pesquisa, que podem ser em um experimento nas ciências, um estudo empírico nas ciências sociais ou observações de fenômenos culturais, até a publicação dos resultados da pesquisa." (PAMPEL et al., 2013, p. 1, tradução nossa).

A disponibilização dos dados científicos para o acesso, usos e reuso, objetivo primeiro dos repositórios de dados, requer um planejamento e gerenciamento eficiente, segundo Simionato (2017) essas ações devem ser iniciadas desde a confecção do *Data Management Plan* (DMP) pelo pesquisador no depósito dos dados durante toda sua execução até sua finalização.

Para que os dados possam ser disponibilizados à comunidade científica se faz necessário apresentar requisitos padronizados e validados segundo normas internacionais e assegurar uma infraestrutura adequada para preservação digital.

Nesse sentido, parte desta infraestrutura equivale a composição dos metadados por meio de seus padrões. Assim, a representação e o armazenamento dos dados é uma preocupação recorrente do profissional que tem a informação como seu objeto de trabalho.

Há grande relação entre dados e metadados. Suas definições podem ser tratadas como muito similares. Jeffery et al. (2014, não paginado, tradução nossa) explicam que a percepção pode variar a partir do contexto da análise. Por exemplo, para um pesquisa que faz uma busca em uma base de dados, ele utiliza metadados para descobrir um livro ou artigo, já para quem gerencia o sistema, a base de dados pode ser utilizada como dados para analisar as coleções por assunto, por editora, por ano etc.

Entre os principais padrões de metadados para descrição de objetos digitais está o *Dublin Core* (DC). O DC surgiu em 1995, com uma proposta de a partir de um núcleo reduzido, representar qualquer documento na *Web*. (ARAKAKI, ALVES, SANTOS, 2018).

Nesse contexto, o armazenamento, representação são fundamentais para recuperação e acesso à informação. Rodrigues et al. (2010) aponta que os repositórios de dados provêem serviços dirigidos à quem deposita e aos provedores do sistema que são como uma extensão dos repositórios digitais, porém afirma que: "Nos repositórios de dados pode-se ir além desta visão de repositórios

de objetos, uma vez que cada conjunto de dados possui características próprias e por isso pode requerer um tratamento diferenciado" (RODRIGUES et al., 2010, p. 23).

Foram criados diversos tipos de repositórios de dados, como os repositórios temáticos que buscam reunir dados de uma área do conhecimento. Há uma tendência ainda, para os repositórios de dados institucionais, quando as Instituições de Ensino disponibilizam os dados oriundos de pesquisa para comunidade e ainda há possibilidade de repositórios relacionados à grupos de pesquisa conforme explorado por Vidotti et al. (2017).

Sales e Sayão (2015, p. 28) ao discorrerem sobre a implantação de repositórios de dados e curadoria digital de dados pontuam que "[...] as exigências sobre o nível de descrição e de atribuição de metadados devem ser identificadas desde o começo de seu projeto e revistas ao longo de vida dos seus dados". Os autores ainda complementam afirmando que essa é a essência da curadoria dos dados, ou seja, os metadados deverão assegurar a disponibilização dos dados para seu uso e reuso.

Nesse cenário, o crescimento do número de repositórios de dados nos leva a seguinte indagação: quais são os padrões de metadados utilizados na representação dos dados científicos disponibilizados nos repositórios de dados na América Latina?

Santos e Alves (2013) ao discorrerem sobre os padrões de metadados apontam que suas estruturas padronizadas representam um conteúdo informacional intencionam sua recuperação e acesso, ou seja, os padrões de metadados são "[...] um conjunto estruturado, padronizado, codificado e pré-determinado 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).

Nesse cenário, esse trabalho tem como objetivo analisar a formalização dos padrões de metadados utilizados para a descrição do conjunto de dados de pesquisa no âmbito dos repositórios de dados da América Latina registrados no R*egistry of Research Data Repositories* (re3data).

2 PROCEDIMENTOS METODOLÓGICOS

O trabalho apresenta os resultados de uma abordagem qualitativa, referente à representação da informação nos repositórios de dados científicos. Para isso, optou-se pela realização de uma pesquisa exploratória e documental da literatura, nacional e internacional, sobre a temática e da descrição de conjunto de dados em repositórios de dados.

Durante a fase de identificação e verificação dos resultados, optou-se pela cobertura geográfica da América Latina. A base escolhida para a busca dos repositórios foi a *re3data.org* (disponível em: https://www.re3data.org/) sendo os resultados recuperados até o dia 04 de agosto de 2018.

A *re3data.org* configura-se como um registro global de repositórios de dados de pesquisa que abrange repositórios de dados de pesquisa de diferentes disciplinas acadêmicas. O filtro utilizado foi a cobertura geográfica, no caso a América Latina, e os resultados apresentaram 29 repositórios de dados, que serão abordados no próximo tópico.

3 RESULTADOS

Dos 29 repositórios identificados, foram localizados onze (11) repositórios no México, oito (8) repositórios no Brasil, Argentina, Colômbia e Panamá foram localizados dois (2) repositórios cada e Chile, El Salvador e Peru foi localizado um (1) repositório cada. Importante relatar que até o dia 4 de agosto de 2018, nos países: Bolívia, Costa Rica, Cuba, Equador, Guatemala, Haiti, Honduras, Nicarágua, Paraguai, República Dominicana, Uruguai e Venezuela não foram encontrados nenhum repositório cadastrado na *re3data.org*.

Durante o levantamento, foi observado que cinco (5) repositórios utilizam o software *Dspace*, três (3) repositórios utilizam *software DataVerse*, dois (2) repositórios utilizam *Drupal,* um (1) repositório utiliza o *Metacat*, e um repositório informou que usa a denominação de "outro", mas não especificou o *software* utilizado. Entre os repositórios analisados, um considerou o *MySQL* como sistema utilizado e 16 repositórios apontaram que o sistema é desconhecido no contexto do *re3data* ou não informaram. O uso de um *software* adequado para organização e tratamento dos dados influencia diretamente nas possibilidades de recuperação, armazenamento, representação da informação e o estabelecimento de padrões de metadados.

Nesse contexto os repositórios da América Latina, os repositórios: *Repositorio Institucional USIL* (Peru), *Repositorio Institucional UCASAL* (Argentina), Base de Dados Científicos da Universidade Federal do Paraná (Brasil), CEDAP *Research Data Repository - research data* (Brasil) adotam o *Dublin Core*. Destaca-se que sua composição básica possui 15 elementos que são opcionais e repetíveis, além da possibilidade de expandir a quantidade de elementos e refinamentos.

Em contrapartida, há diversos outros padrões específicos para descrição de dados no contexto dos repositórios de dados. Conforme o levantamento realizado, três (3) repositórios (CIAT *Dataverse* - Colômbia; IBICT *Dataverse Network* - Brasil; e CIMMYT *Research Data & Software Repository Network* - Mexico) utilizam um padrão internacional para descrição de dados científicos, o *Data Documentation Initiative* (DDI). Os metadados do DDI permitem representar o conjunto de dados e gerenciar diferentes estágios no ciclo de vida dos dados de pesquisa, como conceituação, coleta, processamento, distribuição, descoberta e arquivamento. (DATA DOCUMENTATION INITIATIVE, 2018).

O padrão *Darwin Core* é utilizado no Portal de datos de Biodiversidad (Argentina). O *Darwin Core* está estruturado com base no *Dublin Core*, XML e RDF Schema e inclui um glossário de termos tendo como objetivo facilitar o compartilhamento de informações sobre diversidade biológica. (DARWIN CORE TASK GROUP, 2014).

No contexto dos dados ecológicos, o *PPBio Data Repository* (Brasil) adotou o *Ecological Metadata Language* (EML), que é uma especificação de metadados que está na versão 2.1.1, implementado como uma série de tipos de documentos XML que podem ser usados de maneira modular e extensível. (KNOWLEDGE NETWORK FOR BIOCOMPLEXITY, 2015).

Para descrição de dados geoespaciais, o repositório *Integrated Taxonomic Information System* (ITIS) do México optou por utilizar o *Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata* (FGDC/CSDGM) que estabelece e implementa orientações para descrição do conteúdo, da qualidade e da transferência de dados geoespaciais. Entre os tipos de materiais

estão mapas, arquivos de Sistemas de Informações Geográficas (GIS), imagens e outros recursos. (FEDERAL GEOGRAPHIC DATA COMMITTEE, 2014).

Para o intercâmbio de dados no âmbito da área da astronomia o *Flexible Image Transport System* (FITS) foi adotado pelo repositório *Gran Telescopio CANARIAS Public Archive*, do México. O FITS é um padrão recomendado pela NASA e pela *International Astronomical Union* e usado para o transporte, análise e armazenamento de arquivos de conjuntos de dados científicos, pode apresentar matrizes multidimensionais como espectros 1D, imagens 2D, 3D e mais os cubos de dados. (NASA, 2017).

Entre os repositórios analisados, a ISO 19115 que define o esquema para descrever informações e serviços geográficos por meio de metadados é utilizado no *International Ocean Discovery Program* (Brasil). O padrão da ISO fornece informações sobre a identificação, a extensão, a qualidade, os aspectos espaciais e temporais, o conteúdo, a referência espacial, a representação, a distribuição e outras propriedades dos dados e serviços geográficos digitais. (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2014).

O *Access to Biological Collection Data* (ABCD) é utilizado pelo repositório internacional GLOBE, com abrangência no Brasil. O ABCD é um padrão que abrange a descrição de dados de espécimes de coleções de história natural e de espécies em geral. (BOTANIC GARDEN AND BOTANICAL MUSEUM BERLIN-DAHLEM, 2016).

Diante o levantamento, foi observado que mais da metade dos repositórios analisados, dezesseis (16), não informaram se utilizam algum padrão de metadados ou se optaram pelos metadados já identificados como descritores do *software* aderido.

4 CONSIDERAÇÕES FINAIS

Durante o levantamento realizado, constatou-se que grande parte dos repositórios da América Latina não utilizam ou não informaram os padrões de metadados adotados, isto é, como já apresentado nos resultados 16 repositórios não referenciam à padrões.

O padrão de metadados mais adotado entre os 29 repositórios de dados é o *Dublin Core*, em razão da utilização do *software Dspace* que já faz parte da instalação da plataforma.

Em compensação, alguns repositórios utilizam metadados específicos da área do repositório, possibilitando uma descrição mais exaustiva das informações além de permitir a interoperabilidade entre sistemas.

Outro ponto de destaque, é que os princípios de descrição dos conjuntos de dados são fundamentados nos mesmos princípios de descrição de outros tipos de recursos informacionais, sendo que propriedades de clareza, precisão, lógica e integridade devem ser equalizadas para todos os tipos de recursos, diferenciando-se apenas, em especificidades relativas à forma e formato.

O reconhecimento para a importância dos dados científicos, bem como o seu compartilhamento em repositórios vêm se potencializando com a criação de novas plataformas, *softwares* para o gerenciamento e outras possibilidades para a descrição de conjuntos de dados, entretanto, ainda carece de pesquisas e desenvolvimento.

REFERÊNCIAS

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

ARAKAKI, F. A.; ALVES, R.C.V.; SANTOS, P.L.V.A. da C. Dublin Core: state of art (1995 to 2015). **Informação & Sociedade**: Est., João Pessoa, v. 28, n. 2, p. 7-20, maio/ago. 2018. Disponível em: <http://www.periodicos.ufpb.br/ojs2/index.php/ies/article/view/38012/pdf>. Acesso em: 19 nov. 2018.

BOTANIC GARDEN AND BOTANICAL MUSEUM BERLIN-DAHLEM. **Access to Biological Collection Data schema**. 2016. Disponível em: <http://www.biocase.org/products/schema_repository/>. Acesso em: 6 ago. 2018.

DARWIN CORE TASK GROUP. **Darwin Core**. 2014. Disponível em: <http://rs.tdwg.org/dwc/index.htm>. Acesso em: 6 ago. 2018.

DATA DOCUMENTATION INITIATIVE. **Welcome to the Data Documentation Initiative**. 2018. Disponível em: <http://www.ddialliance.org/>. Acesso em: 6 ago. 2018.

FEDERAL GEOGRAPHIC DATA COMMITTEE. **Geospatial Metadata**. 2014. Disponível em: <https://www.fgdc.gov/metadata>. Acesso em: 6 ago. 2018.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 19115-1:2014** - Geographic information -- Metadata -- Part 1: Fundamentals. Disponível em: <https://www.iso.org/standard/53798.html>. Acesso em: 6 ago. 2018.

JEFFERY, K. et al. A 3-Layer Model for Metadata. INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATION, 13., Portugal, **Anais**… DCMI, EUA. 2014. Disponível em: <http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/199/199>. Acesso em: 17 out. 2015.

KNOWLEDGE NETWORK FOR BIOCOMPLEXITY. **Ecological Metadata Language (EML)**. 2015. Disponível em: <https://knb.ecoinformatics.org/#external//emlparser/docs/index.html>. Acesso em: 6 ago. 2018.

NASA. **FITS Support Office**. 2017. Disponível em: <https://fits.gsfc.nasa.gov/>. Acesso em: 6 ago. 2018.

PAMPEL, H. et al. Making research data repositories visible: The re3data. org registry. **PloS one**, v. 8, n. 11, p. e78080, 2013. Disponível em: <http://bit.ly/2vHTqGA>. Acesso em: 07 out. 2018.

RODRIGUES, E. et al. **Os repositórios de dados científicos:** estado da arte. 2010. Disponível em: <http://projeto.rcaap.pt/index.php?option=com_remository&Itemid=2&func=startdown&id=271&lang=pt>. Acesso em: 15 set. 2018.

SAYÃO L. F.; SALES, L. F. Guia de gestão de dados de pesquisa para bibliotecários e pesquisadore. Rio de Janeiro : CNEN/IEN, 2015. Disponível em:< http://carpedien.ien.gov.br/bitstream/ien/1624/1/GUIA_DE_DADOS_DE_PESQUISA.pdf> Acesso em: 08 mar. 2017.

SIMIONATO, A. C. Mapeamento dos metadados para dados científicos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18. 2017, Marília. **Anais…**: Marília: UNESP, 2015. Disponível em <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/view/563> : Acesso em: 02 de out. 2018.

VIDOTTI, S. A. B. G. et al. Repositório de dados de pesquisa para grupo de pesquisa: um projeto piloto. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18, 2017, Marília. **Anais...** Marilia: UNESP, 2017.Disponível em: <http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/viewFile/388/932>. Acesso em: 17 set. 2018