GESTÃO DE DADOS DA BIODIVERSIDADE: aplicação do padrão de metadados Darwin Core

Biodiversity data management: applying the Darwin Core metadata standard

Filipi Miranda Soares[1], Raíssa Yuri Hamanaka[2], Benildes C. Moreira dos Santos Maculan[3]

(1) Escola de Ciência da Informação, Universidade federal de Minas Gerais, filipivgp2011@gmail.com
(2) Escola de Ciência da Informação, Universidade federal de Minas Gerais, raissa0201@gmail.com
(3) Escola de Ciência da Informação, Universidade federal de Minas Gerais, benildes@gmail.com

Resumo:
No contexto da gestão de dados de pesquisa e da necessidade de padronização do armazenamento e tratamento dos mesmos para posterior reutilização, torna-se imprescindível o uso de metadados para a representação dos dados. Devido a importância do uso de metadados para a recuperação de documentos, o estudo objetivou analisar a aplicação do padrão de metadados Darwin Core (DwC) em um registro de ocorrência do Portal da Biodiversidade. Para avaliar a aplicação do DwC pelo Portal da Biodiversidade, foi feita uma correlação entre os metadados do DwC com os metadados do registro de ocorrência do repositório, buscando-se entender o significado de cada metadado e se eles correspondem de fato aos metadados do DwC, uma vez que a documentação do Portal da Biodiversidade apresenta o DwC como padrão de metadados adotado. Dos 34 campos do registro de ocorrência analisados, a maioria tinha correspondência com os campos do padrão de metadados. Houve casos de duplicação de campos no repositório da biodiversidade, da criação de campos no repositório que não existiam no DwC e da existência de campo inadequado no repositório. A partir do exemplo do Portal da Biodiversidade é possível delinear a importância da curadoria digital, ao procurar agrupar conjuntos de dados semelhantes e estruturá-los de forma a possibilitar o reuso dos mesmos.
Palavras-chave: Gestão de dados de pesquisa; Metadados; Darwin Core.

Abstract:
In the context of the management of research data and the need to standardize the storage and treatment of them for later reuse, it is essential to use metadata to represent the data. Due to the importance of using metadata for document retrieval, the study aimed to analyze the application of the Darwin Core (DwC) metadata standard in a Portal da Biodiversidade occurrence record. To evaluate the application of the DwC by the Portal da Biodiversidade, a correlation was made between the metadata of the DwC with the metadata of the record of occurrence of the repository, trying to understand the meaning of each metadata and if they correspond in fact to the metadata of the DwC, since the documentation of the Portal da Biodiversidade presents DwC as the metadata standard adopted. Of the 34 occurrence record fields analyzed, most corresponded to the metadata standard fields. There have been cases of duplication of fields in the biodiversity repository, the creation of fields in the repository that did not exist in the DwC and the existence of an inadequate field in the repository. Based on the example of the Biodiversity Portal, it is possible to delineate the importance of digital curatorship by seeking to group similar data sets and structure them in a way that allows them to be reused.
Keywords: Research data management; RDM; Metadata; Darwin Core.

1 INTRODUÇÃO

No atual contexto do fenômeno *big data*, em que são gerados grandes volumes de dados não-estruturados em meio digital, torna-se de extrema importância a questão da organização do dado para sua posterior recuperação.

Inseridos nesse contexto estão os pesquisadores, produzindo um volume grande de dados científicos, o que impacta na definição das atuais fontes de informação (para além do periódico científico) e no compartilhamento de dados, e, em consequência, na comunicação científica. Dessa forma, surgem desafios na gestão dos dados de pesquisa disponíveis em rede (SAYÃO; SALES, 2012).

Como solução para esses desafios tem origem o conceito de curadoria digital que "além de reduzir a duplicação de esforços na criação de dados de pesquisa [...] reforça o valor de longo prazo dos dados existentes quando os tornam disponíveis para a

reutilização em novas pesquisas" (SAYÃO; SALES, 2012, p. 184).

Para garantir o reuso dos dados de pesquisa por pesquisadores, é necessário que os sistemas computacionais consigam acessá-los. Assim, tanto dados quanto algoritmos devem ser criados segundo os princípios de reusabilidade, encontrabilidade, acessibilidade e interoperabilidade (FAIR, 2016). Isso pode ser problematizado pelas seguintes questões: "se você disponibilizar seus dados para um cientista ou pesquisador que não esteve envolvido com seu projeto, seriam capazes de entendê-lo?"; "eles conseguiriam utilizar os dados efetiva e apropriadamente?"; "como criar, organizar, gerir, descrever, preservar e compartilhar dados, efetivamente?" (STRASSER et al., 2012, p. 1, tradução nossa).

A efetiva gestão de dados de pesquisa e sua posterior reutilização por demais pesquisadores, depende de um conjunto de práticas durante a coleta, o processamento e a análise dos mesmos. Essas práticas compõem o ciclo de vida dos dados (CVD), que é formado pelos seguintes elementos: planejar, coletar, assegurar, descrever, preservar, descobrir, integrar e analisar (STRASSER et al., 2012). Para Sant'ana (2016), cada etapa ou elemento do CVD possui características em comum: privacidade, integração, qualidade, direito autoral, disseminação e preservação. Sayão e Sales (2012) dividem o CVD em ações: para todo o ciclo de vida (descrição e representação da informação, planejamento da preservação, participação e monitoramento e curadoria e preservação); sequenciais, que se repetem em cada etapa do ciclo (conceitualização, criação ou recebimento, avaliação e seleção, arquivamento, preservação, armazenamento, acesso, uso e reuso e transformação) e ocasionais (eliminação, reavaliação e migração).

O foco deste estudo foi o elemento "descrição" do CVD, definido como "dados que são descritos com precisão e minuciosamente usando os padrões de metadados apropriados" (STRASSER et al., 2012, p. 3, tradução nossa).

Conforme Costa (2017), há algumas iniciativas brasileiras em relação à gestão de dados de pesquisa, e dentre elas se destacam: o Programa FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) de Pesquisa em e-Science; o Portal da Biodiversidade; a Infraestrutura Nacional de Dados Espaciais no Brasil (INDE); as medidas de incentivo ao acesso aberto do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e o desenvolvimento do repositório de dados do Instituto de Energia Nuclear.

Para explicitar o elemento descrição (e seus fatores) do CVD, analisaram-se os metadados de um registro de ocorrência do repositório de dados do Portal da Biodiversidade em comparação com o padrão de metadados Darwin Core (DwC). DwC é um padrão internacional com base em *taxon*, composto por um conjunto de elementos (propriedades, atributos ou conceitos) desenvolvidos para padronizar o compartilhamento de informações sobre diversidade biológica.

2 OBJETIVO

Analisar a aplicação do padrão de metadados Darwin Core (DwC) em um registro de ocorrência do Portal da Biodiversidade[1].

3 PROCEDIMENTOS METODOLÓGICOS

O universo de pesquisa é todos os registros no Portal da Biodiversidade e como amostra foi selecionado o registro de ocorrência da espécie *Rhinella granulosa* (Spix, 1824), leigamente conhecida como sapo da caatinga. O procedimento adotado foi correlacionar os metadados do DwC com o significado atribuído a cada campo do registro e avaliar o seu uso. Assim, as categorias de análise selecionadas para análise foram os metadados adotados para descrição do registro de ocorrência, que foram correlacionadas com o conjunto de metadados do DwC.

[1] Disponível em: <https://portaldabiodiversidade.icmbio.gov.br/portal>. Acesso em: 30 set. 2018.

4 RESULTADOS

A análise determinou que o registro selecionado possui 34 metadados, apresentados em formato de folha de dados (CSV[2]), baixado do Portal da Biodiversidade, que serão apresentados na ordem que aparecem no registro.

Os dois primeiros metadados do registro são <Nome da instituição> e <Sigla da instituição>. Eles representam um mesmo dado e possuem o mesmo equivalente no DwC, que é o metadado <institutionCode>. O primeiro campo de metadado está sem informações e o segundo apresenta a sigla ICMbio. Esses dois campos deveriam ser apenas um campo, pois significam a mesma coisa. Assim, seria necessário apenas definir o formato de entrada de dados: nome por extenso da instituição ou apenas a sigla.

Em seguida, aparece o metadado <Nome da base de dados>, que possui equivalente no DwC como <datasetName> e está preenchido corretamente. O campo seguinte foi denominado <Sigla da base de dados> e não possui equivalente no DwC. A melhor prática sugerida seria unir esses dois campos em apenas um e padronizar a entrada de dados: nome por extenso da base de dados ou apenas sigla.

Depois vem o campo <Responsável pelo registro> que possui o metadado equivalente no DwC <recordedBy>, com valor formatado de acordo com as recomendações do DwC para este campo.

Em sequência, são apresentados dois campos que equivalem a um mesmo campo no DwC, mas foram preenchidos com valores distintos: <Número do registro no portal> e <Número do registro na base de dados>, que equivalem semanticamente ao campo <catalogNumber> no DwC. O primeiro apresenta o valor numérico ‘1067665’ e o segundo o valor “Nº Da Autorização/Licença Sisbio: 44832”.

Após, aparece o campo <Data do registro> que é ambíguo. Não é possível saber se o valor da data que preenche o campo é a data de registro na base de dados ou a data que o evento (ocorrência) aconteceu. No padrão DwC, o metadado para descrever a data de ocorrência é o <eventDate>, presente no registro como <Data do evento>. Quando baixado em formato CSV, o valor do campo <Data do registro> muda de ‘05/02/2017’ para ‘06/02/2017’, a mesma data presente no campo <Data do evento>. Não há um campo de data de registro no DwC. O campo com significado mais próximo seria <dcterms:modified>, que se refere à última data de modificação do registro documental da ocorrência na base de dados. Além disso, o DwC recomenda utilizar esquemas de codificação para preenchimento das datas, como a ISO 8601:2004(E), o que não foi aplicado no Portal da Biodiversidade. O valor do campo <Data do evento> apresenta outro problema: está definido como ‘06/02/2017 a 06/02/2017’. Se o evento aconteceu no dia seis de fevereiro e não foi inserido o intervalo específico de horas em que ele ocorreu, não é necessário inserir a data como intervalo de datas, uma vez que o evento aconteceu em apenas um dia.

O metadado seguinte, <Data de Carência>, não possui equivalente no DwC e não está claro ao que esta data se refere.

Depois, aparece o metadado <Nome científico>, que possui o equivalente <scientificName> no DwC. O valor desse campo no registro é ‘Rhinella granulosa’, que não está de acordo com as recomendações do DwC (DARWIN CORE TASK GROUP, 2015), que orienta indicar o nome científico completo, com autor e data. A melhor prática seria apresentar o nome ‘*Rhinella granulosa* (Spix, 1824)’.

Em seguida vem <nome comum>, que possui o equivalente <vernacularName> no DwC, e está preenchido em conformidade com as melhores práticas recomendadas pelo padrão.

O campo seguinte, <Nome científico na base de dados>, não deveria existir. O nome científico da espécie é um identificador único, ou seja, não pode variar. Logo, é desnecessário que haja dois campos com o mesmo valor no registro de ocorrência.

Logo em seguida aparecem os metadados <Nível taxonômico>, <Número de indivíduos>, <Reino>, <Filo>, <Classe>, <Ordem>, <Família> e <Gênero>, que têm valores que atendem às orientações de

[2] *Comma-separated values* é um formato de arquivo que apresenta dados tabelados.

melhores práticas do DwC, e possuem como equivalentes, respectivamente: <taxonRank>, <individualCount>, <kingdom>, <phylum>, <class>, <order>, <family> e <genus>.

Depois vem o metadado <Espécie>, que repete o mesmo valor de dois outros campos do registro de ocorrência, portanto, é redundante sua presença no registro. A sugestão seria substituir esse metadado por <specificEpithet>, que representa o epíteto[3] da espécie no nome científico.

A seguir aparecem os metadados <Estado de conservação>, <Categoria de Ameaça> e <Status de Sensibilidade>, que ainda não têm equivalência no DwC. Entretanto, uma extensão[4] está sendo desenvolvida para representar informações de estado de conservação.

Após, aparecem metadados com informações de localização: <Localidade>, <País>, <Estado/Província>, <Município>, <Latitude> e <Longitude>, que possuem equivalentes no DwC, respectivamente: <locality>, <country>, <stateProvince>, <municipality>, <decimalLatitude> e <decimalLongitude>, e têm valores conforme as orientações do padrão.

Depois aparecem os campos <Outras informações da localidade> e <Jurisdição>, que não possuem equivalentes no DwC.

Por fim, o metadado <Destino do Material> tem o equivalente no DwC <MaterialSample>.

5 CONSIDERAÇÕES FINAIS

Conforme o objetivo do estudo, foram analisados 34 campos de um registro de ocorrência do Portal da Biodiversidade em correlação com os campos do padrão de metadados DwC. Na maioria das análises, em cerca de 65% dos casos, houve correspondência do campo do registro com o campo do padrão de metadados. Entretanto, houve casos de duplicação de campos no repositório da biodiversidade, valores inadequados, criação de campos que ainda não existem no DwC e existência de campo inadequado no repositório.

O Portal da Biodiversidade reuni registros de ocorrência de nove bases de dados de biodiversidade brasileira, sendo uma iniciativa de gestão de dados que desde 2016 está integrado ao Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr), que está ligado ao Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC).

Essa iniciativa é um exemplo que permite perceber a importância da curadoria digital, que procura agrupar conjuntos de dados semelhantes e estruturá-los de maneira a possibilitar o seu reuso. Para tanto, é necessário manter interoperabilidade entre as diferentes bases, o que pode ser alcançado por meio do uso de padrões de metadados. Considera-se que na implementação do CVD, os metadados são essenciais, pois possibilitam a padronização da representação e/ou codificação dos dados para sua futura recuperação.

Acredita-se que a padronização dos dados coletados e armazenados em repositórios permite a democratização dos dados ao garantir o acesso efetivo aos mesmos.

6 REFERÊNCIAS

COSTA, Maíra Murrieta. **Diretrizes para uma política de gestão de dados científicos no Brasil**. 2017. 288 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, DF, 2017.

DARWIN CORE TASK GROUP. Biodiversity Information Standards. Darwin Core. 2015. Available from: <http://rs.tdwg.org/dwc/>. Accessed on: 8 Sept. 2018.

FAIR principles for data stewardship. **Nature Genetics**, London, v. 48, n. 4, p. 343, Apr. 2016.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. **Inf. Inf.**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016.

SAYÃO, Luis Fernando; SALES, Luana Faria. Curadoria digital: um novo patamar para preservação de dados digitais de

[3] Designa uma espécie ou subespécie diferente dentro de um mesmo gênero.

[4] Disponível em: <https://tools.gbif.org/dwca-validator/extension.do?id=http://purl.org/plic/terms/3.2.1/ThreatStatus>. Acesso em: 25 set. 2018.

pesquisa. **Inf. & Soc.: Est.**, João Pessoa, v. 22, n. 3, p. 179-191, set./dez. 2012.

STRASSER, Carly et al. Primer on data management: what you always wanted to know: but were afraid to ask. **DataONE**, Oakland, CA, p. 1-11, fev. 2012.