GESTÃO DE DADOS DE BIOTECNOLOGIA MARINHA

Luiz Gadelha[a]

[a] Laboratório Nacional de Computação Científica

RESUMO

Um banco de dados pode ser definido como uma coleção organizada de dados. Na pesquisa científica, o uso de bancos de dados é crescente, como pode ser observado em projetos de diversas áreas, como por exemplo, na biologia. Os bancos de dados também estão cada vez mais presentes na rotina dos pesquisadores, em particular nas ômicas. Este capítulo descreve alguns dos principais bancos de dados de biodiversidade, por ex. o *Brazilian Marine Biodiversity Database* (BaMBa), desenvolvido para manter grandes conjuntos de dados do ambiente marinho. Os conjuntos de dados curados obtidos a partir de estudos holísticos integrados, que compreendem parâmetros físico-químicos, -ômicas, microbiologia, pesquisas bentônicas e de peixes, podem ser publicados no banco de dados, possibilitando acesso *online* rápido pela academia, agências regulatórias e a indústria.

1 INTRODUÇÃO

O tamanho sem precedentes da população humana, associado às suas atividades econômicas, tem um impacto cada vez maior nos ambientes globais (Newbold et al., 2015). Em muitos países, isso despertou a preocupação de governos sobre o desequilíbrio entre o consumo de recursos por essas atividades e a capacidade dos ecossistemas de prover recursos. Esse desequilíbrio resultou, por exemplo, em perda de cobertura florestal em muitos lugares, na extinção de espécies e na diminuição da disponibilidade de água potável. Os seres humanos dependem de *serviços ecossistêmicos* em várias atividades. Esses serviços, como alimentos e água, resultam de processos que ocorrem dentro desses ecossistemas. Vários estudos mostram que há uma forte relação entre as atividades humanas, as mudanças globais, a biodiversidade e os processos e serviços ecossistêmicos. Chapin et al. (2000) observaram que as variáveis da biodiversidade, como o número de espécies presentes, o número de indivíduos de cada espécie e quais espécies estão presentes, bem como os tipos de interações (por exemplo, tróficas, competitivas, mutualistas) que ocorrem entre essas espécies, determinam as características intrínsecas de espécies, que são expressas por genes ou afetadas pelo ambiente, que influenciam os processos ecossistêmicos. Os autores também observam que mudanças globais, muitas vezes desencadeadas por seres humanos, como as espécies invasoras, o aumento do dióxido de carbono atmosférico e a mudança no uso da terra, podem alterar significativamente essas variáveis da biodiversidade e, consequentemente, a expressão de características intrínsecas de espécies. Isso, por sua vez, afeta os processos ecossistêmicos e seus serviços resultantes, que podem ter impactos negativos no desenvolvimento humano. As mudanças nesses serviços ecossistêmicos que são devidas a mudanças na biodiversidade podem às vezes ser não-lineares e abruptas, o que pode representar um risco significativo para os seres humanos. Conclusões semelhantes foram alcançadas em outros levantamentos sobre a relação entre a biodiversidade, o funcionamento de ecossistemas e os serviços ecossistêmicos (Cardinale et al., 2012; Hooper et al., 2012). Cardinale et al. (2012) observam que após a extinção de uma espécie, as mudanças resultantes nos processos ecológicos dependem fortemente de quais características intrínsecas dessa espécie foram eliminadas. Hooper et al. (2012) observam que a perda de biodiversidade é tão significativa para mudanças nos ecossistemas quanto os efeitos diretos de mudanças globais, como o aumento no dióxido de carbono na atmosfera e a diminuição da camada de ozônio. Na Bacia Amazônica, que possui uma parcela considerável da biodiversidade do planeta, estudos demonstram que os incêndios e mudanças no uso da terra podem afetar o armazenamento de carbono, a precipitação e os padrões de descarga dos rios. Isso, por sua vez, afeta os serviços ecossistêmicos cruciais para a população local, como a produção de alimentos, a qualidade do ar e a água potável.

Um grande esforço para resolver o problema foi iniciado em 1992, durante a Cúpula da Terra no Rio de Janeiro, com a assinatura da Convenção sobre a Diversidade Biológica (CBD)[1], um

[1] http://www.cbd.int

tratado internacional juridicamente vinculante. Seus principais objetivos são a conservação da biodiversidade, incluindo ecossistemas, espécies e recursos genéticos, e seu uso sustentável e justo. Os países signatários são instados a elaborar e executar uma estratégia para a conservação da biodiversidade, conhecida como *Estratégia e Plano de Ação Nacional de Biodiversidade* (NBSAP), e a implementar mecanismos para monitorar e avaliar a implementação dessa estratégia. Devem comunicar periodicamente os progressos na implementação de seus NBSAPs. O *Plano Estratégico para a Biodiversidade 2011-2020* define as ações a serem tomadas pelos países para alcançar um conjunto de vinte metas até 2020, conhecidas como *Metas de Biodiversidade de Aichi*. Deve-se observar que a Assembleia Geral das Nações Unidas declarou 2011-2020 a Década das Nações Unidas para a Biodiversidade. Em 2012, foi criada a Plataforma Intergovernamental sobre Biodiversidade e Serviços Ecossistêmicos (IPBES) para permitir uma cooperação mais estreita entre cientistas e tomadores de decisão governamentais na avaliação do estado da biodiversidade e dos serviços ecossistêmicos e as suas relações.

Balmford et al. (2005) observam que, para cumprir as metas de conservação da biodiversidade, é essencial tornar os indicadores e o conhecimento abertamente disponíveis para os tomadores de decisão, de forma que eles possam efetivamente utilizá-los. A Rede de Observação da Biodiversidade do Grupo de Observação da Terra (GEO BON) propôs um conjunto de 22 Variáveis Essenciais da Biodiversidade (EBVs) (Pereira et al., 2013) que permitiriam monitorar e avaliar mudanças na biodiversidade. O desenvolvimento e a implantação de mecanismos para produzir esses indicadores dependem do acesso a dados confiáveis de pesquisas de campo, sensores automatizados, coleções biológicas, dados moleculares e literatura acadêmica histórica. A transformação desses dados brutos em dados sintetizados que sejam adequados ao uso requer diversas etapas de refinamento. Deve-se avaliar sua qualidade (Chapman, 2005), observando sua precisão taxonômica, geográfica e temporal. Em muitos casos, a cobertura geográfica dos dados é limitada, exigindo, por exemplo, o uso de modelagem de distribuição de espécies (MDE) (Phillips; Anderson; Schapire, 2006) para estimar a probabilidade de uma determinada espécie ocorrer em alguma região geográfica. As metodologias e técnicas utilizadas para gerenciar e analisar esses dados compreendem uma área muitas vezes chamada *Informática na Biodiversidade* (Hardisty et al., 2013; Hobern et al., 2013; La Salle et al., 2016). Guralnick e Hill (2009), por exemplo, propõem o conceito de um mapa mundial que registraria os padrões globais de biodiversidade e como eles mudam ao longo do tempo, derivado de várias fontes, como sensoriamento remoto, literatura sobre biodiversidade, coleções biológicas e bases de dados de sequências de DNA. A partir desse mapa, várias análises, tais como riqueza de espécies e distribuição, podem ser atualizadas periodicamente usando dados mais recentes.

Neste capítulo, apresenta-se uma visão geral desta área de pesquisa, cobrindo seus principais conceitos, práticas e alguns dos desafios existentes.

2 O QUE SÃO BANCOS DE DADOS

Um banco de dados (Garcia-Molina; Ullman; Widom, 2009) pode ser definido como uma

coleção organizada de dados. Bancos de dados são utilizados em diversas áreas, desde sítios na *web*, como Google e Amazon, a grandes empresas, para manter informações sobre seus negócios. Em Ciência da Computação, a área de pesquisa de Bancos de Dados torna-se mais intensa nos anos 1960, a partir da proposição de bancos de dados onde os dados eram organizados em forma de rede (grafos) ou em forma de árvore (hierárquico). O modelo de dados CODASYL (*Conference/Committee on Data Systems Languages*), baseado em redes, foi bastante utilizado na época, assim como o sistema de banco de dados hierárquico IMS, da IBM. Um dos trabalhos mais influentes e com maior impacto na área de bancos de dados, por Edgar Codd, foi o que apresentou o modelo relacional de bancos de dados (Codd, 1971). Tal modelo permitiu que bancos de dados fossem manipulados com alto grau de abstração, evitando que seus administradores tivessem que se preocupar com detalhes de baixo nível, como estratégias de armazenamento de dados em arquivos. Atualmente é o modelo de dados mais utilizado em sistemas de gerenciamento de bancos de dados comerciais e de *software* livre, a exemplo do Oracle Database, Microsoft SQL Server, PostgreSQL e MySQL.

Na pesquisa científica, o uso de bancos de dados é crescente, como pode ser observado em projetos de diversas áreas. Na Astronomia, por exemplo, diversos levantamentos de objetos celestes foram conduzidos, resultando em bancos de dados dos projetos *Sloan Digital Sky Survey* (SDSS), *Dark Energy Survey* (DES), *Large Synoptic Survey Telescope* (LSST). Na Biologia, também estão cada vez mais presentes na rotina dos pesquisadores, em particular na genômica, com o *GenBank, DNA Data Bank of Japan* (DDBJ), e o banco de dados do *European Molecular Biology Laboratory* (EMBL), e na biodiversidade, com o *Global Biodiversity Information Facility* (GBIF). Atualmente, bancos de dados são ferramentas essenciais, por permitir que dados científicos sejam armazenados de forma consolidada e persistente e disponibilizados para comunidades de cientistas. Neste capítulo, descreveremos alguns dos principais bancos de dados de biodiversidade, como o GBIF e seu nó brasileiro, o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr). Tais bancos de dados, também chamados de repositórios, permitem que sejam consultados dados sobre biodiversidade e que cientistas neles *publiquem* seus próprios conjuntos de dados. Na área de gerenciamento de dados científicos, "publicar" tem o sentido de carregar ou contribuir com um conjunto de dados para algum banco de dados. Na próxima seção, descrevemos as principais etapas do ciclo de vida de dados sobre a biodiversidade.

2.1 CICLO DE VIDA DE DADOS DE BIODIVERSIDADE

De acordo com a abordagem de Michener e Jones (2012), dados sobre a biodiversidade seguem um ciclo de vida composto pelas etapas de planejamento, coleta, certificação, descrição, preservação, descoberta, integração, análise. Vale observar que após a atividade de análise, novos ciclos de gestão de dados de biodiversidade podem ser desencadeados conforme o seu resultado. Tais etapas são ilustradas na Figura 1 e geralmente compõem um *Plano de Gestão de Dados* (PGD) de atividades de pesquisa em biodiversidade. Algumas agências de financiamento de pesquisa em países como os Estados Unidos exigem a apresentação de um PGD em submissões a editais de financiamento.

Figura 1. Etapas do ciclo de vida de dados de biodiversidade (Michener; Jones, 2012).

Pesquisadores, sejam produtores ou consumidores de dados de biodiversidade, provavelmente executarão atividades relacionadas a pelo menos uma dessas etapas. O restante deste capítulo aborda cada uma das etapas do ciclo de vida de dados de biodiversidade, descrevendo as metodologias, ferramentas, recomendações e desafios nelas existentes.

2.2 PLANEJAMENTO

O plano deve conter o que será coletado, quando, por quem e como. Elaborar previamente modelos de planilhas tanto para os metadados como para os dados. É recomendado que haja vocabulários controlados ou ontologias para os termos utilizados durante a coleta de dados.

2.3 TIPOS DE DADOS DE BIODIVERSIDADE

A biodiversidade está relacionada com a variedade de organismos vivos, que pode ser medida de muitas formas e em diferentes escalas, a partir de um registro de um organismo observado em uma localização geográfica em uma determinada data, uma *ocorrência de espécie* (Yessonet al., 2007), à abundância relativa de espécies em uma amostra de água coletada em um sítio de pesquisa ecológica de longa duração (Michener et al., 2011). As ômicas também apresentam muitas oportunidades para explorar a biodiversidade. Por exemplo, dados moleculares de amostras ambientais podem ser analisados em estudos de metagenômica para identificar características funcionais e a classificação taxonômica de organismos presentes. Nesta seção, listamos tipos comuns de dados que são usados para descrever e analisar a biodiversidade.

2.4 OCORRÊNCIAS DE ESPÉCIES

As ocorrências de espécies são um dos tipos de dados mais frequentemente disponíveis sobre biodiversidade. Os atributos principais de uma ocorrência de espécie são dados por: um *táxon*, que é definido como um grupo de uma ou mais populações de organismos que formam uma unidade; uma localização, e uma data de registro da ocorrência. Os registros de ocorrências de espécies são originários de diferentes fontes. As coleções biológicas presentes, por exemplo,

em museus de história natural são uma fonte valiosa para analisar a biodiversidade. Hardisty et al. (2013) observaram que apenas cerca de 10% das coleções de história natural são digitalizadas e que são necessárias ferramentas de informática para acelerar o processo. Kemp (2015) observa que atualmente a maioria das descobertas de novas espécies são baseadas em espécimes armazenados nessas coleções. Esses espécimes geralmente contêm informações detalhadas sobre o local e a data em que foram coletados, as quais podem ser usadas para derivar registros de ocorrência de espécies. Além dos espécimes de coleções biológicas, observações humanas são outra fonte de registros de ocorrências de espécies. Essas observações ocorrem, por exemplo, durante expedições de campo ou até mesmo através de iniciativas de ciência cidadã, como eBird (http://www.ebird.org) e iNaturalist (http://www.inaturalist.org).

2.5 LISTAS DE ESPÉCIES

Frequentemente são realizados levantamentos em uma região geográfica, como um país ou um parque nacional, para determinar quais espécies estão presentes. Essas pesquisas normalmente resultam em uma lista de táxons, comumente chamada de lista de espécies. Também podem ser restritos a um determinado reino ou bioma. Forzza et al. (2012), por exemplo, descrevem como a *Lista da Flora Brasileira* foi elaborada e publicada em 2010. O esforço envolveu a agregação de informações sobre *vouchers* registrados em sistemas de informação de herbários e contou com taxonomistas para revisá-los. O *Catálogo da Vida* (http://www.catalogueoflife.org) agrega mais de 100 listas oficiais de espécies e contém informação sobre cerca de 1,6 milhões de espécies. O *Taxonomy Database* do NCBI (https://www.ncbi.nlm.nih.gov/taxonomy), por exemplo, contém a classificação e nomenclatura dos organismos que estão armazenados em bancos de dados públicos de sequências, como o *GenBank* (https://www.ncbi.nlm.nih.gov/genbank/) e o ENA (http://www.ebi.ac.uk/ena).

2.6 DADOS DE AMOSTRAGENS E OBSERVACIONAIS

Os dados baseados em amostras são coletados em eventos, que podem ocorrer apenas uma vez ou periodicamente repetidos. Envolvem tipicamente dados ecológicos e possuem grande amplitude e heterogeneidade de temas. Podem envolver, por exemplo, levantamentos populacionais e medições abióticas em diversas escalas temporais e espaciais em transectos, grades e parcelas (Magnusson et al., 2013). Um caso típico em que são coletados esses tipos de dados são os *Projetos Ecológicos de Longa Duração* (PELDs)(Michener et al., 2011). Em função da heterogeneidade dos dados ecológicos, ainda não há um vocabulário controlado que seja amplamente utilizado. Algumas iniciativas nessa direção incluem ontologias como a ENVO (*Environment Ontology*), a OBOE (*Extensible Observation Ontology*) e a BCO (*Biological Collections Ontology*) (Walls et al., 2014). As ferramentas mais comuns para publicação de dados ecológicos recorrem a metadados para descrever os conjuntos de dados tabulares que os compõem. Tais metadados permitem que informações gerais como a identificação do proprietário de um conjunto de dados e suas coberturas geográficas, temporais e taxonômicas sejam registradas, facilitando a interpretação pelos usuários. Os metadados permitem também descrever textualmente

o significado de cada coluna de um conjunto de dados tabular. Mais adiante, neste capítulo, descreve-se a *Ecological Metadata Language* (EML), um padrão de metadados para conjuntos de dados ecológicos.

2.7 ÔMICAS

A análise de sequências de DNA, RNA e de proteínas tem diversas aplicações no estudo da biodiversidade. As sequências genéticas obtidas de amostras ambientais contendo comunidades de organismos, i.e. metagenomas (Robbins et al., 2012), por exemplo, fornecem informações importantes para analisar as características taxonômicas e funcionais dos seus membros. As atividades de taxonomia podem ser auxiliadas pela análise de sequências de DNA (Tautz et al., 2003). O projeto *Barcoding of Life* (Ratnasingham; Hebert, 2007), por exemplo, analisa regiões pequenas e padronizadas dos genes para auxiliar a identificação de espécies. Alguns sistemas, como o VoSeq (Peña; Malm, 2012), permitem ligar *vouchers* presentes em coleções biológicas a sequências de DNA presentes em bancos de dados genômicos. Guralnick e Hill (2009) observam que a biodiversidade pode ser mais precisamente quantificada, quando comparada à simples contagem do número de espécies, pelo grau de relacionamento filogenético entre as espécies. Como exemplos, eles avaliam a prioridade de conservação de aves norte-americanas usando sua distinção filogenética e risco de extinção e analisam a dispersão do vírus influenza A, também usando a análise filogenética.

2.8 LITERATURA CIENTÍFICA

Uma vasta quantidade de informações a respeito da biodiversidade está contida na literatura acadêmica. Sínteses de expedições de campo frequentemente estão disponíveis somente em artigos científicos, cujos dados relacionados às amostragens, coletas e as respectivas análises estão propagados para bancos de dados sobre a biodiversidade. Algumas iniciativas, como a *Biodiversity Heritage Library* (BHL)(Gwinn; Rinaldo, 2009), visam utilizar tecnologias, como o reconhecimento ótico de caracteres, para extrair essas informações de artigos científicos e disponibilizá-las em bancos de dados públicos.

2.9 IMAGENS E VÍDEOS

Expedições de campo para realização de amostragens frequentemente envolvem a produção de imagens e vídeos que auxiliam a análise dos sítios estudados. Nas seções seguintes, descrevemos o *Audubon Core*[2], um vocabulário controlado para a descrição de recursos multimídia associados a dados de amostragens e de ocorrências de espécies.

2.10 COLETA

[2] https://terms.tdwg.org/wiki/Audubon_Core

Dados de biodiversidade podem ser coletados de diversas formas: redes de biossensores, expedições de campo, observações feitas por cientistas cidadãos, entre outras. No processo de coleta, é importante que sejam utilizados identificadores únicos para projeto, evento de amostragem, área de amostragem e protocolo utilizado (Stocks; Stout; Shank, 2016). Esses identificadores permitirão adiante que os dados coletados sejam armazenados em bancos de dados de forma consistente. Sempre que possível, os termos devem seguir algum vocabulário controlado ou ontologia, como a BCO-DMO (http://www.bco-dmo.org).

2.11 CERTIFICAÇÃO

Soberón e Peterson (2004) listam problemas comuns em relação a dados de biodiversidade. Espécimes de coleções biológicas, de onde é extraída parte considerável dos dados de ocorrências de espécies, podem possuir identificações incorretas ou com taxonomia desatualizada. A taxonomia biológica está em constante mudança para acomodar novos conhecimentos sobre espécies. É possível também que haja georreferenciamento incorreto, decorrente de erros de anotação ou imprecisão de instrumentos. Em registros antigos, pela indisponibilidade de mecanismos para avaliação precisa de localização, é comum encontrar somente descrições textuais do local onde um espécime foi coletado.

Diversas ferramentas podem ser utilizadas para reduzir ou eliminar erros de identificação de espécies. Vários catálogos oficiais de espécies oferecem serviços acessíveis pela web para consulta de táxons, como o *Catalogue of Life*[3], *World Register of Marine Species*[4] e a *Lista de Espécies da Flora do Brasil*[5]. É recomendado que as identificações realizadas em dados coletados relativos a observações de espécies sejam validadas em algum desses catálogos oficiais. Boa parte desses catálogos é também acessível por meio de interfaces de programação (APIs) disponíveis na web, permitindo a automação desse tipo de verificação com scripts ou aplicações.

Com relação a problemas de georreferenciamento, Guralnick (2009) menciona a importância de determinar a incerteza do georreferenciamento dos registros de ocorrências e seu impacto na escala na qual estudos podem ser realizados. Ferramentas como o BioGeomancer (Guralnick, 2006) e Geolocate[6] tentam inferir quais são as coordenadas geográficas de uma ocorrência de espécie a partir de uma descrição textual da sua localização. Otegui e Guralnick (2016) propõem uma API acessível pela *web*[7] que realiza verificações simples de consistência nos registros, como coordenadas com valor zero, campo país discordante das coordenadas, coordenadas invertidas.

2.12 DESCRIÇÃO

Na etapa de descrição, um conjunto de metadados é elaborado para descrever um con-

[3] http://www.catalogueoflife.org/
[4] http://www.marinespecies.org/
[5] http://floradobrasil.jbrj.gov.br/
[6] http://www.museum.tulane.edu/geolocate/
[7] http://api-geospatial.vertnet-portal.appspot.com/geospatial

junto de dados sobre biodiversidade. Esses metadados são essenciais para que um usuário que obtenha um conjunto de dados consiga interpretá-lo. Nesta seção, descrevemos os padrões, práticas e recomendações para documentação de dados sobre a biodiversidade.

2.13 *ECOLOGICAL METADATA LANGUAGE* (EML)

A *Ecological Metadata Language* (EML) (Fegraus, 2005) é um padrão de metadados desenvolvido originalmente para a descrição de dados ecológicos. Atualmente é também usada para documentar conjuntos de dados relativos a observações de espécies. O padrão possui diversos perfis com seus respectivos campos que podem ser utilizados para definir os atributos de um conjunto de dados. Um perfil de *descrição científica* contém campos como o criador do conjunto de dados (*creator*), sua cobertura geográfica (*geographicCoverage*), temporal (*temporalCoverage*) e taxonômica (*taxonomicCoverage*), e protocolo de amostragem (*sampling*). Esse perfil é utilizado para definir atributos do conjunto de dados como um todo.

O perfil de *representação de dados*, através da entidade *data Table*, permite descrever os atributos de um conjunto de dados tabular. Podem ser definidos os tipos de dados de tais atributos, como datas e valores numéricos, assim como suas restrições, como valores mínimos e máximos. Utilizados em conjunto, os perfis de descrição científica e de representação de dados podem fornecer uma documentação de boa qualidade para um conjunto de dados, facilitando consideravelmente sua interpretação pelos usuários.

Em termos de implementação, um conjunto de metadados em EML é elaborado com a linguagem XML, ilustrada na Figura 2.

```
<eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.1.1"
        xmlns:dc="http://purl.org/dc/terms/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.1 http://rs.gbif.org/schema/eml-gbif-profile/1.1/eml.xsd"
        packageId="1edcfe6d-da55-4d59-b30e-468cd21f8b0b/v2.6" system="http://gbif.org" scope="system"
        xml:lang="eng">

  <dataset>
    <alternateIdentifier>1edcfe6d-da55-4d59-b30e-468cd21f8b0b</alternateIdentifier>
    <alternateIdentifier>http://ipt.sibbr.gov.br/sibbr/resource?r=bamba_fish_seamount</alternateIdentifier>
    <title xml:lang="eng">Fish biodiversity of the Vitória-Trindade Seamount Chain, Southwestern Atlantic: an updated database</title>
    <creator>
      <individualName>
        <givenName>Hudson</givenName>
        <surName>Pinheiro</surName>
      </individualName>
      <organizationName>University of California Santa Cruz</organizationName>
      <positionName>Researcher</positionName>
    </creator>
    <pubDate>
      2016-12-08
    </pubDate>
    <language>eng</language>
    <abstract>
      <para>Primary data was acquired during three scientific diving expeditions to the VTC seamounts and islands, in 2009 (12-26 March)
and 2011 (3-26 February and 1-18 April). These expeditions covered the photic and upper mesophotic zones (17-120 m depth) of the two
islands and eight seamounts: Almirante Saldanha, Vitoria, Eclaireur, Jaseur, (Columbia Bank in [35]), Davis, Dogaressa and Columbia seamo
unts. Sampling included visual, video and photo records, as well as collection of voucher specimens by divers (hand nets and spear-guns i
n April 2011) using technical open-circuit SCUBA or closed-circuit rebreathers (Megalodon®) with mixed-gases (TRIMIX and EAN). Primary
data from fishery surveys (surface longline, bottom longline, midwater trawling and angling activities; see [31"34,69,105]) were incorporate
d in the database. Fishery sampling was performed over eight volcanic mounts (Vitoria, Eclaireur, Besnard, Montague, Jaseur, Davis, Dogare
ssa, Columbia and Trindade) by the REVIZEE and to a much lesser extent TAMAR/ICMBio monitoring assessments.</para>
    </abstract>
    <keywordSet>
      <keyword>Occurrence</keyword>
      <keywordThesaurus>GBIF Dataset Type Vocabulary: http://rs.gbif.org/vocabulary/gbif/dataset_type.xml</keywordThesaurus>
    </keywordSet>
    <keywordSet>
      <keyword>Observation</keyword>
      <keywordThesaurus>GBIF Dataset Subtype Vocabulary: http://rs.gbif.org/vocabulary/gbif/dataset_subtype.xml</keywordThesaurus>
    </keywordSet>
```

Figura 2. Metadados especificados conforme o padrão EML.

Normalmente, os bancos de dados de biodiversidade disponibilizam ferramentas para edição e elaboração de metadados no padrão EML de forma mais amigável, através de uma interface gráfica. O repositório de dados observacionais DataONE[8] (Michener et al., 2012), por exemplo, permite que usuários forneçam metadados através de uma ferramenta gráfica chamada Morpho (Higgins; Berkley; Jones, 2002), ilustrada na Figura 3. O mesmo repositório possui também uma *interface web* chamada Metacat (Berkley et al., 2001), que permite a carga de dados ecológicos tabulares em formato livre documentados com o padrão EML. O padrão EML é utilizado também para descrição de conjuntos de dados formatados sobre ocorrências de espécies e amostragens, como será descrito na seção seguinte.

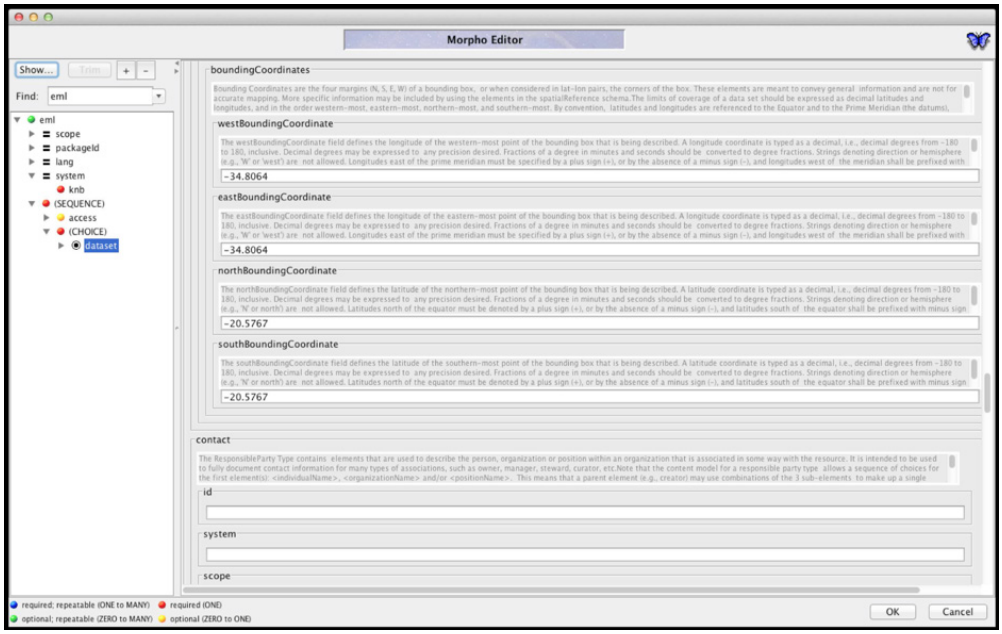

Figura 3. Interface de usuário da ferramenta Morpho para inserção de metadados.

2.14 PRESERVAÇÃO

Na etapa de preservação, os conjuntos de dados sobre biodiversidade são publicados em algum banco de dados, como o DataONE e o GBIF, onde estarão disponíveis para a comunidade científica. Tais bancos de dados adotam práticas de curadoria e gestão dos dados, visando sua preservação e disponibilidade em longo prazo. Existem diversos procedimentos possíveis para a publicação. Neste capítulo, serão descritos padrões e procedimentos para carga de um conjunto de dados em um banco de dados de biodiversidade. Será descrito também o fluxo de publicação dos principais repositórios atuais.

2.15 *DARWIN CORE*

Darwin Core (Wieczorek et al., 2012) é um padrão para representação de dados de bio-

[8] https://www.dataone.org/

diversidade que visa facilitar o compartilhamento dos mesmos. O padrão é composto por uma lista de termos relativos à biodiversidade e suas respectivas definições. As discussões, evolução e manutenção do *Darwin Core* (DwC) são realizadas pelo TDWG (*Biodiversity Information Standards*) (http://www.tdwg.org), uma associação para o desenvolvimento e promoção de padrões para o registro e intercâmbio de dados sobre a biodiversidade. O DwC surgiu como um perfil de termos no sistema *Species Analyst*, em 1998, desenvolvido pela Universidade do Kansas para gerenciamento de coleções biológicas. Em 2002 foi adotado para o intercâmbio de informações no *Mammal Networked Information System* (MaNIS), um sistema distribuído composto por várias instituições mantenedoras de coleções biológicas de mamíferos. Em 2009, foi iniciado o processo de padronização do DwC, que foi ratificado em outubro do mesmo ano na reunião anual do TDWG. O DwC se baseia no padrão *Dublin Core* (http://dublincore.org/), aproveitando seus termos para a descrição de recursos, como tipo (*type*), modificado por (*modified*) e licença (*license*), e complementando-os com termos específicos de biodiversidade, como número de catálogo (*catalogNumber*) e nome científico (*scientificName*).

Os termos do vocabulário do DwC são organizados da seguinte forma: as classes indicam as categorias ou entidades definidas no padrão. São exemplos de classes: evento (*Event*), localidade (*Location*) e táxon (*Taxon*). Cada classe possui um conjunto de propriedades, que são seus atributos. Por exemplo, a classe Location tem atributos como *country* e *decimalLatitude*. Finalmente, *valores* podem ser atribuídos às propriedades, como "Chile", -33.61 para as propriedades country e decimalLatitude, respectivamente. Vale observar que é recomendado que, sempre que possível, os valores sejam provenientes de algum vocabulário controlado, no caso de valores textuais, ou de algum padrão de formatação, no caso de valores numéricos ou temporais. Por exemplo, nomes de espécies provenientes de alguma lista reconhecida de espécies, como o *Catalogue of Life* (http://www.catalogueoflife.org). A Tabela 1 ilustra a representação de dados de ocorrências de espécies com o DwC. Esses registros são provenientes de um conjunto de dados publicado através do *Brazilian Marine Biodiversity Database* (BaMBa) no GBIF (http://www.gbif.org/dataset/1edcfe6d-da55-4d59-b30e-468cd21f8b0b).

Tabela 1. Ocorrências de espécies representadas com o DwC.

id	eventDate	decimalLatitude	decimalLongitude	scientificName
6	2002-08-01	-20.805828	-37.761231	*Alectis ciliaris* (Bloch, 1787)
118	2002-08-01	-22.382222	-37.587500	*Balistes vetula* (Linnaeus, 1758)
141	2002-08-01	-19.848744	-38.134635	*Caranx crysos* (Mitchill, 1815)
507	2002-08-01	-20.525417	-29.310350	*Thunnus obesus* (Lowe, 1839)

Normalmente, um conjunto de dados no formato DwC vem acompanhado por metadados, que são definidos no padrão EML (*Ecological Metadata Language*) (Fegraus et al., 2005). No EML, são encontrados campos como o título, autores, cobertura geográfica e temporal do conjunto de dados, que auxiliam usuários a interpretarconjuntos de dados formatados no padrão DwC.

A exemplo de bancos de dados relacionais, conjuntos de dados que seguem o formato

DwC podem conter múltiplas tabelas que se relacionam através de propriedades que são comuns a todas elas. Tal organização permite, por exemplo, que dados de amostragens sejam expressos também nesse padrão. As Tabelas 2 e 3 ilustram esse tipo de organização de dados para representar amostragens de espécies. A Tabela 2 contém os eventos de amostragem, quatro no total. A coluna *eventId* contém um identificador para cada evento. As demais colunas descrevem a data do evento, a latitude e a longitude, respectivamente. A Tabela 3 contém contagens de organismos para cada evento. A coluna *eventId* descreve a que evento da Tabela 1 as contagens se referem. Por exemplo, as primeiras duas linhas da tabela se referem ao evento que possui identificador 1, que está associado a uma amostragem realizada em 18 de março de 2009.

Tabela 2. Eventos de amostragem representados com o DwC.

eventId	*eventDate*	*decimalLatitude*	*decimalLongitude*
1	2009-03-18	-20.51	-38.07
2	2009-03-18	-20.57	-34.80
3	2011-02-11	-20.50	-25.35
4	2011-02-11	-20.47	-34.80

Tabela 3. Ocorrências de espécies relacionadas aos eventos da Tabela 1.

eventId	*organismQuantity*	*scientificName*
1	2	*Alectis ciliaris* (Bloch, 1787)
1	5	*Balistes vetula* (Linnaeus, 1758)
2	1	*Caranx crysos* (Mitchill, 1815)
...	...	...

Tabela 4. Referências sobre o DwC.

Página oficial do padrão DwC:	http://rs.tdwg.org/dwc
Guia de referência rápida do DwC:	http://rs.tdwg.org/dwc/terms
Termos DwC (Português)	http://www.sibbr.gov.br/areas/index.php?area=publicar&subarea=termos-dwc

A seguir, são descritos os procedimentos para publicação de conjuntos de dados de biodiversidade nos principais repositórios e sistemas de informação da área. É através desses sistemas que os dados são disseminados e preservados.

3 SISTEMAS DE INFORMAÇÃO EM BIODIVERSIDADE

3.1 SISTEMA DE INFORMAÇÃO SOBRE A BIODIVERSIDADE BRASILEIRA (SiBBr)

O Sistema de Informação sobre a Biodiversidade Brasileira[9] (SiBBr) (Gadelha et al., 2014) é uma iniciativa do Ministério da Ciência, Tecnologia, Inovação e Comunicações que disponibiliza uma infraestrutura para o gerenciamento de dados de biodiversidade do Brasil. Todos os dados publicados no SiBBr são propagados para a rede global, GBIF. Assim, o SiBBr atua como o nó brasileiro dessa rede. A descrição arquitetural vigente do SiBBr, aqui apresentada, lista componentes já implementados e disponíveis na página *web* do sistema. O SiBBr é organizado de forma modular, onde boa parte da funcionalidade está disponível através de serviços *web*. Dessa forma, novas funcionalidades podem ser agregadas gradualmente e os componentes existentes podem ser modificados e adaptados de forma flexível. Os dois principais componentes para gerenciamento de dados são providos pelo *Portal de Espécies e Ocorrências* e pelo *Portal de Dados Ecológicos*. O primeiro é responsável pela coleta, indexação e disseminação de conjuntos de dados, registros de ocorrência e listas de espécies disponibilizadas por diversos publicadores de dados. Atualmente estão publicados mais de 10 milhões de registros.

Esses conjuntos de dados são publicados utilizando o padrão *Darwin Core* (Wieczorek et al., 2012) com metadados no padrão EML (Fegraus et al., 2005). A funcionalidade de publicação de dados é provida pela ferramenta *Integrated Publishing Toolkit* (IPT) (Robertson et. al, 2014), mostrada na Figura 4, que permite que dados disponíveis localmente nas instituições participantes, como, por exemplo, em planilhas ou bancos de dados relacionais, sejam mapeados para o padrão *Darwin Core* e coletados pelo *Portal de Ocorrências e Listas de Espécies*.

As instituições que desejam publicar dados sobre ocorrências e listas de espécies no SiBBr recebem um registro globalmente único que as identifica tanto no escopo do SiBBr quanto do GBIF. Cada instituição pode instalar e cadastrar instâncias do IPT para publicar os dados disponíveis na sua instituição. Dados disponíveis em diversos formatos são mapeados para o padrão *Darwin Core* e os metadados correspondentes são criados no padrão EML. A partir daí, ambos os arquivos são empacotados pelo IPT e transformados em um arquivo único no formato *Darwin Core Archive* (DwC-A). Em seguida, o processo de coleta e indexação é feito pelo *Portal de Ocorrências e Listas de Espécies*. Esta é uma ferramenta *web* que acessa os endereços dos IPTs cadastrados no GBIF, realizando o *download* dos DwC-A dos conjuntos de dados, e faz a extração dos dados, indexando-os à base de dados do SiBBr. Tal processo é ilustrado na Figura 5.

O *Portal de Dados Ecológicos*, baseado no Metacat (Berkley et al., 2001), é responsável pelo recebimento, armazenamento e disseminação de conjuntos de dados de, por exemplo, Pesquisas Ecológicas de Longa Duração (PELD) (Michener et al., 2011). Diferentes abordagens em ecologia, aliadas às tradições de pesquisa distintas, tanto em suas subdisciplinas como em

[9] http://www.sibbr.gov.br

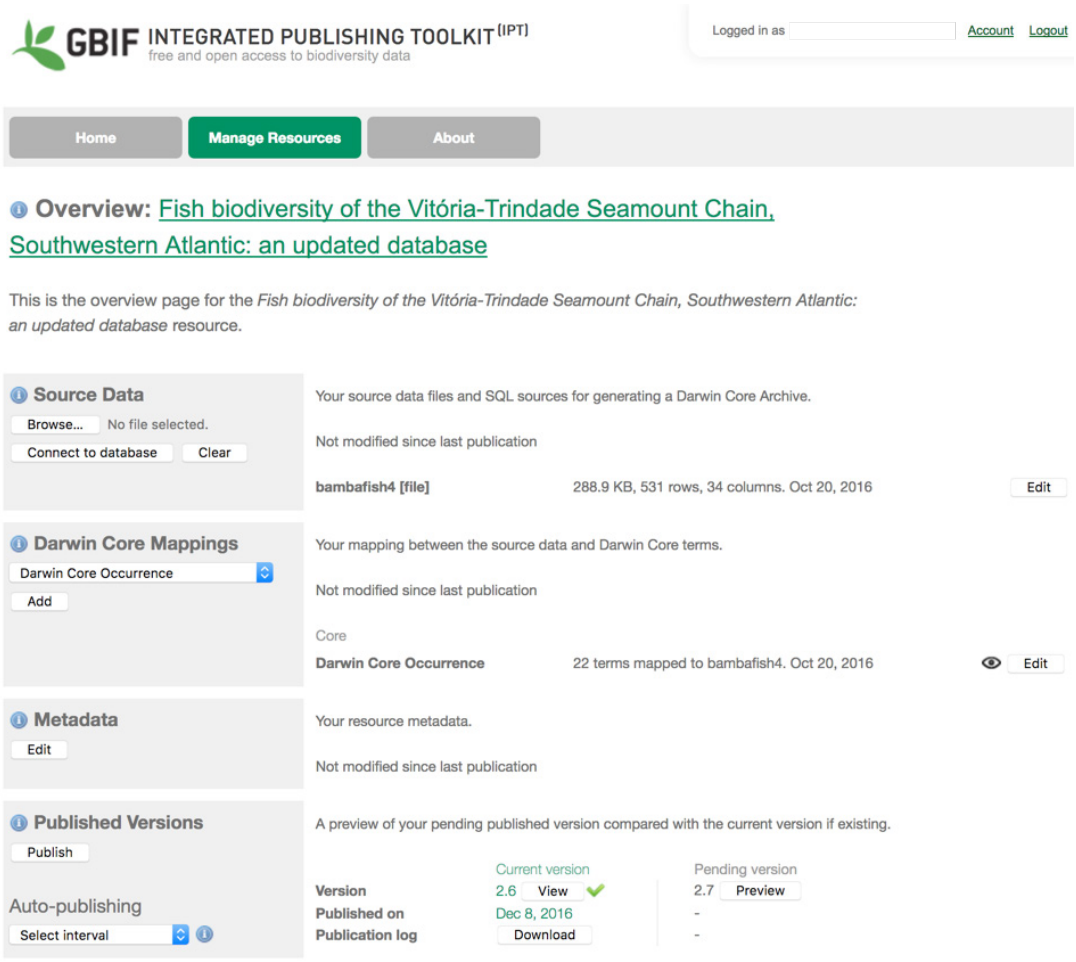

Figura 4. Interface do IPT para publicação de conjuntos de dados de biodiversidade.

áreas afins, levam à produção de dados altamente heterogêneos. Tais dados podem ser, entre outros, contagens de indivíduos, medidas de variáveis ambientais ou representações de processos ecológicos. As terminologias utilizadas também variam de acordo com a linha de pesquisa, bem como a forma de estruturar os dados digitalmente (Jones et al., 2006). Também foi adotado o padrão de metadados EML para a descrição dos conjuntos de dados ecológicos. Os conjuntos de dados em si, em função da heterogeneidade, são publicados no formato original, através de planilhas ou arquivos textuais com valores separados por vírgula.

3.2 BRAZILIAN MARINE BIODIVERSITY DATABASE (BaMBa)

O *Brazilian Marine Biodiversity Database*[10] (BaMBa) (Meirelles et al., 2015) foi desenvolvido para manter grandes conjuntos de dados do ambiente marinho brasileiro. Essencialmente, qualquer informação ambiental pode ser adicionada ao BaMBa. Os conjuntos de dados certificados obtidos a partir de estudos holísticos integrados, que compreendem parâmetros físico-quí-

[10] https://marinebiodiversity.lncc.br

Figura 5. Processo de publicação de dados com o IPT no SiBBr(Gadelha et al., 2014).

micos, -ômicas, microbiologia, pesquisas bentônicas e de peixes, podem ser publicados no banco de dados, possibilitando políticas e ações científicas, industriais e governamentais a serem realizadas em recursos marinhos. Há um número significativo de bases de dados, no entanto o BaMBa é o único recurso de banco de dados integrado que é ao mesmo tempo apoiado por uma iniciativa governamental e exclusivo para dados marítimos. O BaMBa está ligado ao SiBBr e oferece oportunidades para melhorar a governança dos recursos marinhos e a integração dos cientistas.

O processo de publicação de dados com metadados em EML e dados estruturados em DwC no BaMBa segue o procedimento descrito na seção anterior, utilizando a ferramenta IPT. Os dados do BaMBa são propagados para o SiBBr e o GBIF. Os conjuntos já publicados no BaMBa por esse procedimento compreendem tanto ocorrências de espécies quanto dados de amostragens de sítios ecológicos.

Para os casos em que os usuários queiram fornecer somente os metadados e os dados no seu formato tabular original, como planilhas e arquivos de texto separados por vírgula (CSV), é possível utilizar a ferramenta Metacat (Berkleyet al., 2001). Ela permite preencher os principais campos do EML referentes ao conjunto de dados que se deseja publicar através de um formulário *web*. Em seguida, o conjunto de dados pode ser carregado para o repositório no seu formato original. As Figuras 6 e 7 ilustram o processo. Opcionalmente, a entidade *dataTable* do EML pode ser utilizada para fornecer a descrição dos atributos (colunas) dos dados que sejam tabulares. Esse processo de publicação de dados é mais simples, quando comparado ao IPT, uma vez que os dados não precisam ser reestruturados para o formato DwC. Por outro lado, os conjuntos de dados publicados nesse processo requerem maior esforço de interpretação de usuários que os baixem.

A publicação de dados e sua decorrente preservação contribuem para a comunidade científica como um todo. Os dados podem ser reutilizados por outros cientistas que podem explorá-los por outros pontos de vista. Para o publicador dos dados, também podem ser observados benefícios. Um estudo recente (Piwowar; Vision, 2013) mostra que artigos que disponibilizam os dados utilizados em suas análises em repositórios públicos tendem a ter um maior número de citações.

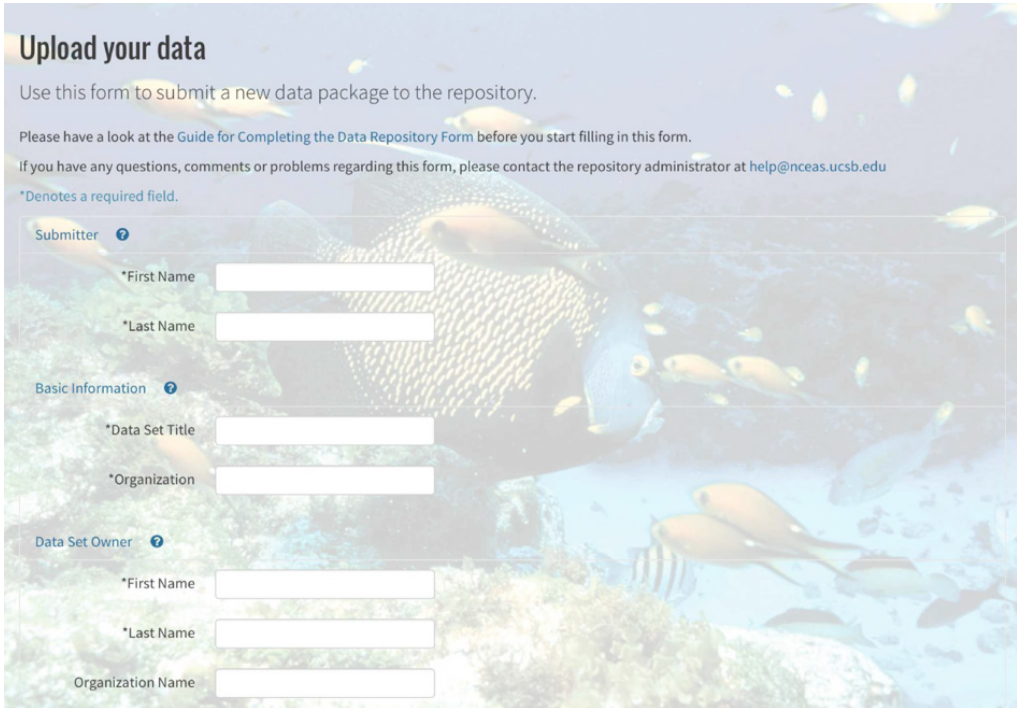

Figura 6. Publicação de dados ecológicos no BaMBa com o Metacat, descrição do conjunto de dados.

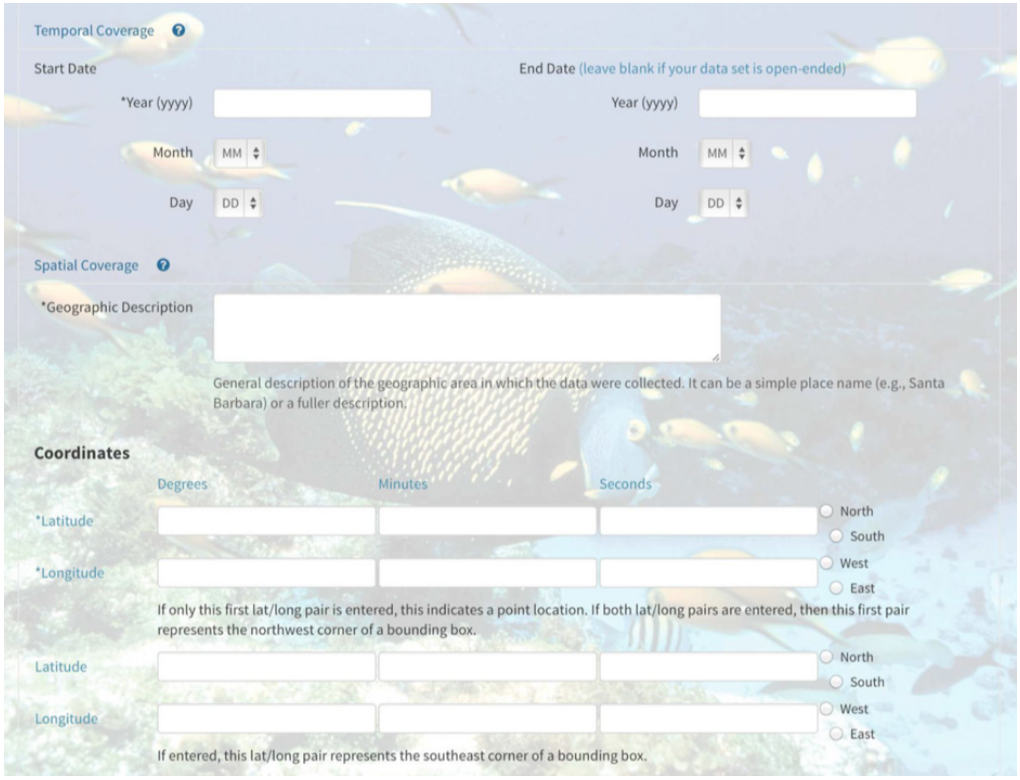

Figura 7. Publicação de dados ecológicos no BaMBa com o Metacat, coberturas temporal e geográfica.

3.3 DESCOBERTA

A busca por dados para a realização de pesquisas de análise e síntese em biodiversidade ainda se apresenta como um desafio. Os avanços mais recentes ocorreram com o surgimento de bancos de dados que agregam conjuntos de dados em escalas globais e nacionais, como o GBIF, DataONE, SiBBe e BaMBa. O uso de padrões de metadados e de publicação de dados permite que instituições mapeiem as representações internas dessas informações para um formato que é claramente especificado e que pode ser consumido e processado de forma automática por máquina. Os bancos de dados agregadores de informações de biodiversidade permitem que os conjuntos de dados sejam buscados de forma geográfica (como na Figura 8), taxonômica e temporal.

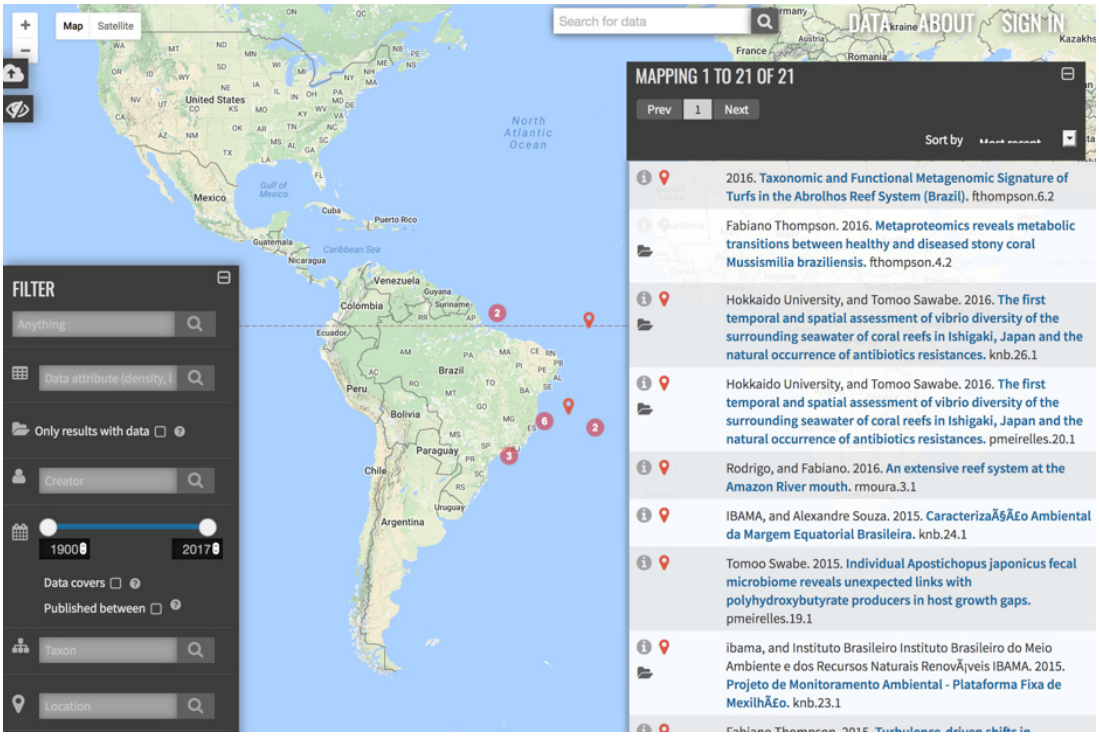

Figura 8. Interface baseada no Metacat do BaMBa para busca de dados ecológicos.

Linguagens e ambientes para análise de dados como o R e o Python já possuem pacotes e bibliotecas que estão integrados aos repositórios e agregadores de dados de biodiversidade. O *rgbif*[11], por exemplo, é um pacote para o R que permite fazer buscas e recuperar registros diretamente do GBIF, sendo o *pygbif*[12] seu análogo para Python.

3.4 INTEGRAÇÃO

Frequentemente cientistas precisam combinar dados provenientes de diferentes fon-

[11] https://cran.r-project.org/web/packages/rgbif/
[12] https://recology.info/2015/11/pygbif/

tes em pesquisas integrativas. Por exemplo, dados físico-químicos podem ser combinados com dados de metagenômica para tentar estabelecer correlações que expliquem algum fenômeno. A atividade de combinar dados provenientes de diferentes fontes é chamada de *integração de dados* e é uma das áreas mais ativas de pesquisa sobre a gerência de dados científicos (Ailamaki; Kantere; Dash, 2010). Os bancos de dados de biodiversidade atuais avançaram ao conseguir estabelecer padrões para os metadados, como o EML, e para os dados, como o DwC. No entanto, estes se limitam a definir vocabulários controlados, consistindo de termos padronizados em cada um dos temas. Uma abordagem mais sofisticada, que envolva não somente a definição de termos, mas também relacionamentos entre estes e regras de inferência, que são chamadas de *ontologias*, são assunto da área de pesquisa de *Web* Semântica. Algumas iniciativas nessa direção na área de biodiversidade e ecologia incluem ontologias como a ENVO, a OBOE e a BCO (Walls et al., 2014). As ontologias permitem cruzar diferentes domínios (*Linked Data*) e realizar consultas semânticas, propiciando uma ferramenta de integração de dados consideravelmente mais poderosa que as atuais.

3.5 ANÁLISE

Os dados científicos estão sendo produzidos a uma taxa de crescimento exponencial por sensores científicos cada vez mais disponíveis. Isso, aliado a sofisticados modelos computacionais que consomem esses dados, exige novas técnicas para gerenciar experimentos computacionais de forma escalável. Esses experimentos são frequentemente compostos por muitas tarefas computacionais que trocam dados através de relações de produção e consumo dos mesmos, e são normalmente especificadas como *workflows científicos* (Ferreira da Silva et al., 2017). Os *sistemas de gerenciamento de workflows científicos* fornecem recursos como tolerância a falhas, execução escalável, gerenciamento escalável de dados, rastreamento de dependências de dados e registro de informações de proveniência, que reduzem consideravelmente a complexidade do gerenciamento do ciclo de vida desses experimentos. Informações de proveniência (Carata et al., 2014), em particular, podem apoiar a análise do resultado de um experimento científico computacional, uma vez que registram o histórico de sua execução. A biodiversidade segue a mesma tendência de rápido aumento na produção de dados. Atualmente, os dados sobre biodiversidade estão sendo integrados em uma escala global, por meio de iniciativas como o GBIF. Aplicações para MDE usam esses conjuntos de dados de biodiversidade, em conjunto com dados ambientais (como climatologia) para prever a distribuição geográfica de uma espécie particular, ou de múltiplas espécies. No SiBBr, por exemplo, rotinas frequentes de análise e síntese são apoiadas por sistemas de gerenciamento de *workflows* científicos (Deelman et al., 2009). Técnicas de computação paralela e distribuída serão utilizadas na execução dessas análises, a exemplo da modelagem de distribuição de espécies (Townsend; Peterson, 2011), através de recursos computacionais de alto desempenho do Sistema Nacional de Processamento de Alto Desempenho[13] (SINAPAD). No SiBBr[14], diferentes estratégias de implementação

[13] http://www.lncc.br/sinapad
[14] https://github.com/sibbr/sdm-workflows

foram realizadas para execução escalável de *workflows* científicos de MDE usando o Swift (Wilde et al., 2011), um sistema de gerenciamento de *workflows* científicos que se concentra na execução paralela e distribuída de tarefas computacionais. Também foi mostrado como o registro de informações de proveniência, conforme suportado pelo Swift com o MTCProv (Gadelha et al., 2012), permite a análise de execução de *workflows* científicos.

4 CONCLUSÃO

Dados de biodiversidade e ecologia são produzidos atualmente a uma taxa exponencial. O GBIF, por exemplo, disponibiliza (em 2017) cerca de 800 milhões de registros de ocorrências de espécies. Estudos integrativos em biodiversidade e ecologia, que necessitam recuperar dados provenientes de diferentes fontes, se beneficiam dessa tendência. Um cientista que adote as práticas, recomendações e ferramentas para gerenciamento de dados de biodiversidade e ecologia beneficiará não apenas à comunidade como um todo, mas também a si próprio. Suas análises tenderão a ter mais impacto em função da melhor transparência e reprodutibilidade da sua pesquisa. Neste capítulo, descrevemos o ciclo de vida pelo qual passam os dados de biodiversidade. Em cada uma das etapas deste ciclo de vida, descrevemos as práticas, ferramentas e recomendações que permitem sua execução. Em particular, na etapa de publicação e preservação dos dados, mostramos o fluxo de publicação dos principais sistemas disponíveis. Diversos desafios ainda existem nessa área, que podem servir de inspiração aos leitores. Pesquisas sobre resolução de entidades, i.e., o mapeamento automático de atributos usados em dados tabulares para termos de vocabulários controlados ou ontologias, podem facilitar consideravelmente o processo de publicação de dados. A migração de vocabulários controlados para ontologias pode desencadear consultas semânticas e inferências automatizadas a respeito de dados de biodiversidade. Finalmente, metodologias para o registro e documentação dos passos executados, em atividades de análise e síntese da biodiversidade, i.e. sua proveniência (Carata et al., 2014), poderiam facilitar a reprodutibilidade e confiabilidade das mesmas.

REFERÊNCIAS

Ailamaki, A.; Kantere, V.; Dash, D. (2010) Managing scientific data. **Communications of the ACM**. 53(6): 68. http://doi.org/10.1145/1743546.1743568

Balmford, A.; Bennun, L.; Brink, B. T.; Cooper, D.; Côte, I.M.; Crane, P.; ...Walther, B.A. (2005) Ecology: The Convention on Biological Diversity's 2010 target. **Science** (New York, N.Y.). 307(5707): 212–3. http://doi.org/10.1126/science.1106281

Berkley, C.; Jones, M.; Bojilova, J.; Higgins, D. (2001) Metacat: a schema-independent XML database system. In: **Proceedings Thirteenth International Conference on Scientific and Statistical Database Mana-**

gement. SSDBM 2001 (pp. 171–179). **IEEE Comput. Soc.** http://doi.org/10.1109/SSDM.2001.938549

Carata, L.; Akoush, S.; Balakrishnan, N.; Bytheway, T.; Sohan, R.; Selter, M.; Hopper, A. (2014) A primer on provenance. **Communications of the ACM**. 57(5): 52–60. http://doi.org/10.1145/2596628

Cardinale, B.J.; Duffy, J.E.; Gonzalez, A.; Hooper, D.U.; Perrings, C.; Venail, P.; ...Naeem, S. (2012) Biodiversity loss and its impact on humanity. **Nature**. 486(7401): 59–67. http://doi.org/10.1038/nature11148

Chapin, F.S.; Zavaleta, E.S.; Eviner, V.T.; Naylor, R.L.; Vitousek, P.M.; Reynolds, H.L.; ... Díaz, S. (2000) Consequences of changing biodiversity. **Nature**. 405(6783): 234–42. http://doi.org/10.1038/35012241

Chapman, A.D. (2005) **Princípios de Qualidade de Dados**. GBIF. http://www.gbif.org/orc/?doc_id=5990.

Codd, E.F. (1971) A relational model for large shared data banks. **Communications of the ACM**. 13(6): 377-387.

Deelman, E.; Gannon, D.; Shields, M.; Taylor, I. (2009). Workflows and e-Science: An overview of workflow system features and capabilities. **Future Generation Computer Systems**. 25(5): 528–540. https://doi.org/10.1016/j.future.2008.06.012

Droege, G.; Barker, K.; Astrin, J.J.; Bartels, P.; Butler, C.; Cantrill, D.; ...Seberg, O. (2014) The Global Genome Biodiversity Network (GGBN) Data Portal. **Nucleic Acids Research**. 42(Database issue): D607-12. http://doi.org/10.1093/nar/gkt928

Fegraus, E.H.; Andelman, S.; Jones, M.B.; Schildhauer, M. (2005) Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. **Bulletin of the Ecological Society of America**. 86(3): 158-168. http://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2

Ferreira da Silva, R.; Filgueira, R.; Pietri, I.; Jiang, M.; Sakellariou, R.; Deelman, E. (2017) A characterization of workflow management systems for extreme-scale applications. **Future Generation Computer Systems**. http://doi.org/10.1016/j.future.2017.02.026

Forzza, R. C., Baumgratz, J. F. A., Bicudo, C. E. M., Canhos, D. A. L., Carvalho, A. A., Coelho, M. A. N., ... Zappi, D. C. (2012). New Brazilian Floristic List Highlights Conservation Challenges. **BioScience**, 62(1), 39-45. http://doi.org/10.1525/bio.2012.62.1.8

Gadelha, L.M.R.; Wilde, M.; Mattoso, M.; Foster, I. (2012) MTCProv: a practical provenance query framework for many-task scientific computing. **Distributed and Parallel Databases**. 30(5-6): 351–370. http://doi.org/10.1007/s10619-012-7104-4

Gadelha, L.; Guimarães, P.; Moura, A.M.; Drucker, D.P.; Dalcin, E.; Gall, G.; ...Leo, W.V. (2014) SiBBr: Uma Infraestrutura para Coleta, Integração e Análise de Dados sobre a Biodiversidade Brasileira. In: **VIII Brazilian e-Science Workshop** (BRESCI 2014). Proc. XXXIV Congress of the Brazilian Computer Society.

Garcia-Molina, H.; Ullman, J.D.; Widom, J. (2009) **Database Systems: The Complete Book** (Second). Pearson Prentice Hall.

Guralnick, R.P.; Wieczorek, J.; Beaman, R.; Hijmans, R.J. (2006) BioGeomancer: automated georeferencing to map the world's biodiversity data. **PLoS Biology**. 4(11): e381. http://doi.org/10.1371/journal.pbio.0040381

Guralnick, R.; Hill, A. (2009) Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. **Bioinformatics** (Oxford, England). 25(4): 421–428. http://doi.org/10.1093/bioinformatics/btn659

Gwinn, N.E., Rinaldo, C. (2009) The Biodiversity Heritage Library: sharing biodiversity literature with the world. **IFLA Journal**. 35(1): 25–34. http://doi.org/10.1177/0340035208102032

Hardisty, A.; Roberts, D.; Addink, W.; Aelterman, B.; Agosti, D.; Amaral-Zettler, L.; ...Young, F. (2013) A decadal view of biodiversity informatics: challenges and priorities. **BMC Ecology**. 13(1): 16. http://doi.org/10.1186/1472-6785-13-16

Higgins, D.; Berkley, C.; Jones, M.B. (2002) Managing heterogeneous ecological data using Morpho.In: Proceedings 14th International Conference on Scientific and Statistical Database Management (pp. 69–76). **IEEE Comput. Soc.** http://doi.org/10.1109/SSDM.2002.1029707

Hobern, D.; Apostolico, A.; Arnaud, E.; Bello, J.C.; Canhos, D.; Dubois, G.; ... Willoughby, S. (2013) **Global Biodiversity Information Outlook - Delivering Biodiversity Knowledge in the Information Age**. Retrieved from http://www.biodiversityinformatics.org/download-gbio-report/

Hooper, D.U.; Adair, E.C.; Cardinale, B.J.; Byrnes, J.E.K.; Hungate, B.A.; Matulich, K.L.; ... O'Connor, M.I. (2012) A global synthesis reveals biodiversity loss as a major driver of ecosystem change. **Nature**. 486(7401): 105–8. http://doi.org/10.1038/nature11118

Kemp, C. (2015) Museums: The endangered dead. **Nature**. 518(7539): 292–294. http://doi.org/10.1038/518292a

La Salle, J.; Williams, K.J.; Moritz, C.; Essl, F.; Dullinger, S.; Rabitsch, W.; ...Moylan, E. (2016) Biodiversity analysis in the digital era. Philosophical Transactions of the Royal Society of London. Series B, **Biological Sciences**. 371(1702): 534–547. http://doi.org/10.1098/rstb.2015.0337

Magnusson, W.; Braga-Neto, R.; Pezzini, F.; Baccaro, F.; Bergallo, H.; Penha, J.; ... Pontes, A.R.M. (2013). **Biodiversity and Integrated Environmental Monitoring**. Áttema Editorial. Retrieved from http://ppbio.inpa.gov.br/sites/default/files/Biodiversidade e monitoramento ambiental integrado.pdf

Meirelles, P.M.; Gadelha Jr., L.M.R.; Francini-Filho, R.B.; Leão, R.M.; Amado-Filho, G.M.; Bastos, A.C.; ... Thompson, F.L. (2015) BaMBa: towards the integrated management of Brazilian marine environmental data. **Database**. http://doi.org/10.1093/database/bav088

Michener, W.K.; Allard, S.; Budden, A.; Cook, R.B.; Douglass, K.; Frame, M.; ...Vieglais, D.A. (2012) Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. **Ecological Informatics**. 11: 5–15. http://doi.org/10.1016/j.ecoinf.2011.08.007

Michener, W.K.; Jones, M.B. (2012) Ecoinformatics: supporting ecology as a data-intensive science. **Trends in Ecology & Evolution**. 27(2): 85–93. http://doi.org/10.1016/j.tree.2011.11.016

Michener, W.K.; Porter, J.; Servilla, M.; Vanderbilt, K. (2011) Long term ecological research and information management. **Ecological Informatics**. 6(1): 13–24. http://doi.org/10.1016/j.ecoinf.2010.11.005

Newbold, T.; Hudson, L.N.; Hill, S.L.L.; Contu, S.; Lysenko, I.; Senior, R.A.; ...Purvis, A. (2015) Global effects of land use on local terrestrial biodiversity. **Nature**. 520(7545): 45–50. http://doi.org/10.1038/nature14324

Otegui, J.; Guralnick, R.P. (2016) The Geospatial Data Quality REST API for Primary Biodiversity Data. **Bioinformatics** (Oxford, England). btw057-. http://doi.org/10.1093/bioinformatics/btw057

Parr, C.S.; Guralnick, R.; Cellinese, N.; Page, R.D.M. (2012) Evolutionary informatics: unifying knowledge about the diversity of life. **Trends in Ecology & Evolution**. 27(2): 94–103. http://doi.org/10.1016/j.tree.2011.11.001

Peña, C.; Malm, T. (2012) VoSeq: A voucher and DNA sequence web application. **PLoS ONE**. 7(6): 1–4. http://doi.org/10.1371/journal.pone.0039071

Pereira, H.M.; Ferrier, S.; Walters, M.; Geller, G.N.; Jongman, R.H.G.; Scholes, R.J.; ... Wegmann, M. (2013) Ecology. Essential biodiversity variables. **Science.** (New York, N.Y.), 339(6117): 277–8. http://doi.org/

10.1126/science.1229931

Phillips, S.J.; Anderson, R.P.; Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions.**Ecological Modelling**, 190(3–4): 231–259. http://doi.org/10.1016/j.ecolmodel.2005.03.026

Piwowar, H.A.; Vision, T.J. (2013) Data reuse and the open data citation advantage. **PeerJ**. 1:e175. http://doi.org/10.7717/peerj.175

Ratnasingham, S.; Hebert, P.D.N. (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). **Molecular Ecology Notes**. 7(3): 355–364. http://doi.org/10.1111/j.1471-8286.2007.01678.x

Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011) Challenges and opportunities of open data in ecology. **Science** (New York, N.Y.). 331(6018): 703–5. http://doi.org/10.1126/science.1197962

Robbins, R.J.; Amaral-Zettler, L.; Bik, H.; Blum, S.; Edwards, J.; Field, D.; ...Wooley, J. (2012) RCN4GSC Workshop Report: Managing Data at the Interface of Biodiversity and (Meta)Genomics, March 2011. **Standards in Genomic Sciences**. 7(1): 159–65. http://doi.org/10.4056/sigs.3156511

Robertson, T.; Döring, M.; Guralnick, R.; Bloom, D.; Wieczorek, J.; Braak, K.; ...Desmet, P. (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. **PLoS ONE**. 9(8): e102623. http://doi.org/10.1371/journal.pone.0102623

Soberón, J.; Peterson, A.T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London. Series B, **Biological Sciences**. 359(1444),: 689–98. http://doi.org/10.1098/rstb.2003.1439

Stocks, K.I.; Stout, N.J.; Shank, T.M. (2016) Information Management Strategies for Deep-Sea Biology. In: Clark, M.R.; Consalvey, M.; Rowden, A.A. (Eds.) **Biological Sampling in the Deep Sea** (pp. 368–385). Wiley Blackwell.

Tautz, D.; Arctander, P.; Minelli, A.; Thomas, R.H.; Vogler, A.P. (2003) A plea for DNA taxonomy. **Trends in Ecology & Evolution**. 18(2): 70–74. http://doi.org/10.1016/S0169-5347(02)00041-1

Peterson, A.T.; Soberón, J.; Pearson, R.G.; Anderson, R.P.; Martínez-Meyer, E.; Nakamura, M.; Araújo, M.B. (2011).**Ecological Niches and Geographic Distributions**. Princeton University Press.

Tuama, E.Ó.; Deck, J.; Dröge, G.; Döring, M.; Field, D.; Kottmann, R.; ...Yilmaz, P. (2012) Meeting Report: Hackathon-Workshop on Darwin Core and MIxS Standards Alignment (February 2012). **Standards in Genomic Sciences**. 7(1): 166–70. http://doi.org/10.4056/sigs.3166513

Walls, R.L.; Guralnick, R.; Deck, J.; Buntzman, A.; Buttigieg, P.L.; Davies, N.; ...Zheng, J. (2014) Meeting report: advancing practical applications of biodiversity ontologies. **Standards in Genomic Sciences**. 9(1): 17. http://doi.org/10.1186/1944-3277-9-17

Wieczorek, J.; Bloom, D.; Guralnick, R.; Blum, S.; Döring, M.; Giovanni, R.; ...Vieglais, D. (2012) Darwin Core: an evolving community-developed biodiversity data standard. **PloS One**. 7(1): e29715. http://doi.org/10.1371/journal.pone.0029715

Wilde, M.; Hategan, M.; Wozniak, J.M.; Clifford, B.; Katz, D.S.; Foster, I. (2011) Swift: A language for distributed parallel scripting. **Parallel Computing**. 37(9): 633–652. http://doi.org/10.1016/j.parco.2011.05.005

Yesson, C.; Brewer, P. W.; Sutton, T.; Caithness, N.; Pahwa, J.S.; Burgess, M.; ...Culham, A. (2007) How global is the global biodiversity information facility? **PloS One**. 2(11): e1124. http://doi.org/10.1371/journal.pone.0001124