A GESTÃO DE DADOS E INFORMAÇÕES SOBRE BIODIVERSIDADE

NO MINISTÉRIO DO MEIO AMBIENTE E INSTITUIÇÕES VINCULADAS

PRÓ ESPÉCIES
Todos contra a extinção

João Monnerat Lanna

A gestão de daados e informações sobre biodiversidade no minitério do meio ambiente e instituições vinculadas

João Monnerat Lanna

Consultor / Biólogo MSC
(CRBIO:70187/04-D)

Contrato NO:
002103-2020

Produto elaborado no âmbito do Projeto Pró-Espécies

Rio de Janeiro-RJ, 15 de junho de 2021

Apresentação

A gestão eficiente de dados e informações sobre biodiversidade é, no âmbito do MMA e suas vinculadas, fundamental para o uso do conhecimento nas tomadas de decisão em conservação e uso sustentável dos recursos naturais. Esforços na compilação e qualificação prévia destes dados e informações beneficiam processos, como por exemplo, a definição de estratégias nacionais e geração de planos de ação para conservação, e gestão de áreas protegidas.

No entanto, é comum o armazenamento de dados em infraestruturas independentes e não integradas, não padronizados, sem metadados e não compartilhados eficientemente. Estes fatores dificultam ou mesmo inviabilizam o uso de dados e informações para dar dão suporte aos processos mencionados.

Este estudo visa avaliar o alinhamento do MMA e suas vinculadas com os princípios e boas práticas de gestão de informação da Informática da Biodiversidade e, complementarmente, aos Princípios FAIR de compartilhamento de dados. Foram consideradas diretrizes técnicas já elaboradas no âmbito do MMA para armazenamento, compartilhamento e publicação de dados e informações de biodiversidade. Direcionam-se assim, ações focadas em melhorias na governança, capacidade e competências na gestão de dados sobre biodiversidade no MMA e suas vinculadas.

Este documento representa um produto de consultoria especializada contratada no âmbito do Projeto GEF Pró-Espécies sobre gestão de dados de biodiversidade brasileira. É também resultado de um esforço conjunto de diversos agentes do MMA, IBAMA, ICMBio e JBRJ, que cederam entrevistas e atuaram na validação e revisão do conteúdo apresentado. Ao final há uma proposta de Plano de Ação para Gestão de Dados com compilação de ações, objetivos e metas, que representa um passo importante para a integração das bases de dados sobre biodiversidade no âmbito do MMA e suas vinculadas, podendo também servir de modelo para outras instituições, públicas e privadas.

Sumário

INTRODUÇÃO 1

A Informática da Biodiversidade e a Interoperabilidade de Sistemas de Informação 2

Padrões de dados e metadados de biodiversidade 4

Princípios FAIR 5

Publicação de dados para pessoas e máquinas 7

- Dados para humanos via repositórios de dados 7
- Dados de biodiversidade via Integrated Publishing Toolkit (IPT) 8
- Dados para máquinas via serviços WEB 9

A OFERTA E INTEGRAÇÃO DE DADOS DE BIODIVERSIDADE NO MMA E VINCULADAS 12

Sistemas de informação sobre biodiversidade do MMA e vinculadas 12 12

- Categorias de dados 17 17
  - Dados taxonômicos 17 17
  - Dados de Ocorrências Biológicas 18 18
  - Dados de Avaliações de Risco de Extinção 20 20
  - Dados de uso e ameaças 21 21
  - Dados sobre Unidades de Conservação 23 23
  - Dados não estruturados 24 24

Integração de dados entre sistemas de informação de biodiversidade do MMA e vinculadas 25

DEMANDAS DO MMA POR DADOS E INFORMAÇÕES SOBRE BIODIVERSIDADE

Estudo de caso – Estratégia Nacional para Conservação de Espécies 27

Potencialidades no uso de dados e informações adequadamente ofertadas 29

Portais de dados de biodiversidade 30

ESTRUTURA DE UM PLANO DE AÇÃO PARA GESTÃO E INTEGRAÇÃO DE DADOS E INFORMAÇÕES DE BIODIVERSIDADE NO MMA E VINCULADAS

Evolução Modular de uma Infraestrutura para Gestão e Integração de Dados e Informações sobre Biodiversidade do MMA e vinculadas 34

CONCLUSÕES

BIBLIOGRAFIA

Glossário

API - application programming interface - Termo genérico que neste contexto é sinônimo de serviço Web;

BI dashboard – Ferramenta de gestão e síntese de informação que é usado para rastreamento de métricas, e outros dados relativos a departamentos ou processos específicos;

EML - Ecological Metadata Language - Padrão de metadados para dados de biodiversidade;

DwC-A – Darwin Core Archive - Padrão de dados desenvolvido para dados de biodiversidade;

IPT – Integrated Publishing Toolkit - Ferramenta de publicação de dados no formato Darwin Core Archive;

Metadados – Conjunto de informações que descrevem um conjunto de dados, correspondem aos “dados sobre os dados”.

XML - É uma recomendação da W3C para gerar linguagens de marcação para necessidades especiais. É um dos subtipos da SGML capaz de descrever diversos tipos de dados. Seu propósito principal é a facilidade de compartilhamento de informações por intermédio da internet.

Introdução

Fontes integradas de dados e informações sobre biodiversidade estão entre as principais e mais urgentes demandas para tomada de decisões em conservação de ecossistemas e preservação de espécies (HOBERN et al., 2019). Atualmente, apesar da grande quantidade de informações disponíveis, a elaboração de estratégias para conservação no âmbito político-institucional ainda demanda um grande esforço de compilação de dados de fontes diversas e despadronizadas.

Sob a gestão do Ministério do Meio Ambiente (MMA) e suas instituições vinculadas, Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA), Instituto Chico Mendes de Conservação da Natureza (ICMBio) e Jardim Botânico do Rio de Janeiro (JBRJ), existe um conjunto heterogêneo de dados e informações sobre a biodiversidade brasileira. Estes incluem dados taxonômicos de flora e fauna, dados de ocorrência de espécies, dados e metadados relativos às avaliações de estado de conservação, Planos de Ação Nacionais para conservação de espécies ameaçadas de extinção (PAN), Planos de Ação Territoriais (PAT), mapas de ocorrência e de modelagem de nicho, dados de ameaças e usos das espécies, imagens, dentre outros. Portanto, a agregação deste conjunto de dados e informações representa um desafio.

No âmbito governamental e institucional, considera-se ainda a iniciativa Governo Aberto e a Política de Dados Abertos do Poder Executivo Federal [1] como importantes conjuntos de regras e compromissos relacionados a temas como a disponibilização de dados governamentais, incluindo dados de pesquisa/ciência.

No 4º Plano de Ação Brasileiro [2], destaca-se o Compromisso 3 [3] presente no tema Inovação e Governo Aberto na Ciência. Este compromisso intitulado "Estabelecer mecanismos de governança de dados científicos para o avanço da Ciência Aberta no Brasil", possui uma série de marcos que tratam do tema Ciência Aberta.

---

1. https://wiki.dados.gov.br/Politica-de-Dados-Abertos.ashx
2. https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/4oplano-de-acao-brasileiro
3. https://wiki.rnp.br/display/OGPBrasil

Estes marcos incluem, por exemplo, a "Implantação de uma rede interinstitucional pela Ciência Aberta", a "Implantação de infraestrutura federada piloto de repositórios de dados de pesquisa" e a "Proposição de padrões de interoperabilidade para repositórios de dados de pesquisa".

Na Política de Dados Abertos do Poder Executivo Federal, destaca-se o trecho do capítulo III do Decreto Nº 8.777 (2016) que trata dos Planos de Dados Abertos (PDA) institucionais: "A implementação da Política de Dados Abertos ocorrerá por meio da execução de Plano de Dados Abertos no âmbito de cada órgão ou entidade da administração pública federal, direta, autárquica e fundacional [...]".

O Decreto Nº 8.777 (2016) explicita que estes PDAs devem, no mínimo, dispor de tópicos envolvendo mecanismos transparentes de priorização na abertura de bases de dados; considerar o potencial de utilização e reutilização dos dados tanto pelo Governo quanto pela sociedade civil; apresentar mecanismos para a promoção, o fomento e o uso eficiente e efetivo das bases de dados pela sociedade e pelo Governo; dentre outros.

Torna-se então evidente a necessidade da adoção de políticas claras por parte de cada instituição sobre sua oferta de dados/informações para a sociedade, por meio da elaboração de um PDA e pela execução das políticas contidas nos compromissos de Ciência Aberta dos quais o Brasil é participante e explicitadas no 4º Plano de Ação Brasileiro do Governo Aberto.

A INFORMÁTICA DA BIODIVERSIDADE E A INTEROPERABILIDADE DE SISTEMAS DE INFORMAÇÃO

O conjunto de dados primários sobre biodiversidade envolve fontes como taxonomia, biogeografia e ecologia. A sintetização destes dados brutos em informação útil requer uma série de passos que seguem metodologia de gestão e análise, que constituem uma área normalmente chamada de informática da biodiversidade (GADELHA et al., 2020).

Existem três grupos de atores inter-relacionados envolvidos neste tipo de dados: (1) Provedores de Dados, como museus de história natural, herbários, sociedades de ciência cidadã, sistemas de informação que são alimentados com dados primários, etc; (2) Agregadores/Publicadores de Dados, iniciativas que fornecem dados vindos de diversos provedores; (3) Usuários de Dados, incluindo cientistas, analistas, tomadores de decisão e público em geral (ANDERSON et al., 2020; GRAHAM et al., 2004). A Figura 1, retirada de (SILVA et al., 2015), ilustra a relação entre estes grupos.

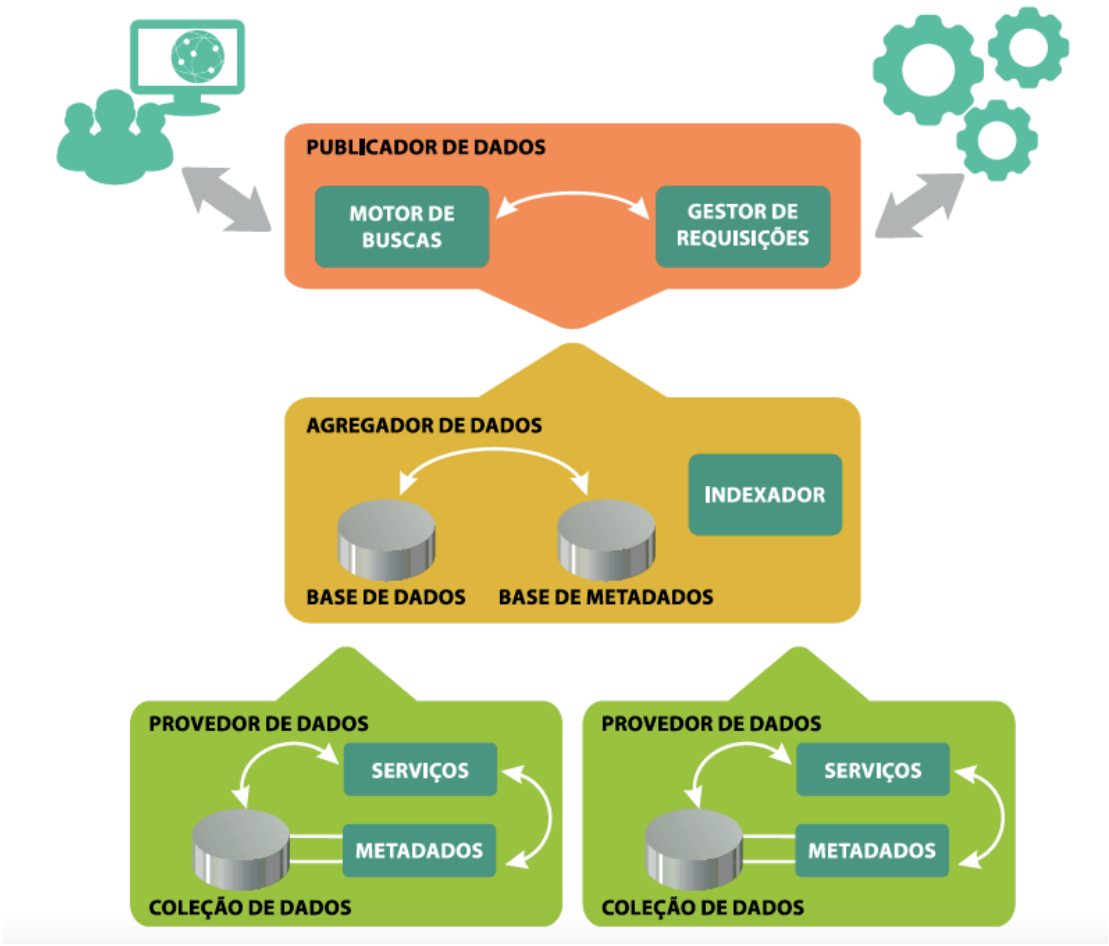

Figura 1. Arquitetura para interoperabilidade de dados de biodiversidade (Silva, 2013, apud Silva et al. 2015)

Integrados por padrões de dados como o Darwin Core Archive (DwC-A), enormes volumes de dados brutos existem hoje na web. O Global Biodiversity Information Facility (GBIF) corresponde ao maior e mais diverso agregador, atualmente com aproximadamente 1,9 bilhões de registros de ocorrência de aproximadamente 1600 instituições.

https://www.gbif.org/

PADRÕES DE DADOS E METADADOS DE BIODIVERSIDADE

A importância do uso de padrões e protocolos que sejam adequados ao contexto e sigam semântica correspondente, se baseia na redução de risco de erros, na interoperabilidade, escalabilidade de sistemas e consumo de dados por terceiros. No contexto de biodiversidade, os padrões de dados são primariamente definidos pela organização Biodiversity Information Standards (TDWG).

Neste tópico serão descritos, de maneira sucinta, os principais padrões de dados e metadados utilizados na gestão de dados e informação sobre biodiversidade. Optou-se por não se aprofundar neste tema e indicar o livro desenvolvido no âmbito do Ministério do Meio Ambiente, Diretrizes para a Integração de Dados de Biodiversidade (SILVA et al., 2015), que possui um excelente nível de detalhes e descrições detalhadas sobre o assunto.

No escopo deste trabalho, vamos considerar os dois padrões de dados, Darwin Core Archive (DwC-A) (WIECZOREK et al., 2012) e Dublin Core (WEIBEL; KOCH, 2000), e os padrões de metadados Ecological Metadata Language (EML) (VITOUSEK et al., 2005) e padrão INDE para dados e metadados geoespaciais .

Objetivamente, o DwC-A é o padrão utilizado para padronização de dados como registros de ocorrências e listas taxonômicas. É composto por um arquivo central, ou núcleo do recurso, onde as informações básicas do táxon ou da ocorrência são descritas, e arquivos chamados extensões que abrigam informações complementares. Este conjunto de arquivos se relacionam pelo identificador único de cada táxon ou ocorrência, presente no arquivo central (Figura 2). O padrão de metadados EML, no escopo deste diagnóstico, é utilizado para descrever os "dados sobre os dados" que foram padronizados em DwC-A, ou seja, funciona para descrever os conjuntos de dados de ocorrência e listas taxonômicas, organizando as informações relacionadas a estes, funcionando como um documento anexo aos recursos de informação.

---

https://www.inde.gov.br/CatalogoMetadados

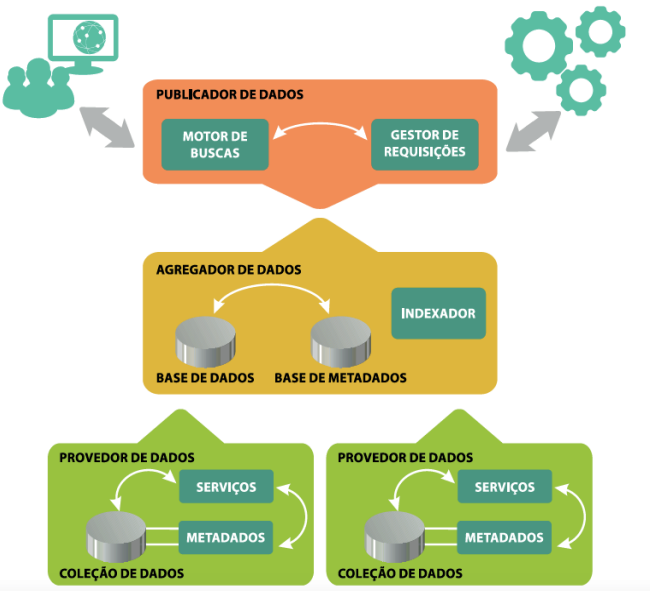

Figura 1. Arquitetura para interoperabilidade de dados de biodiversidade (Silva, 2013, apud Silva et al. 2015)

Integrados por padrões de dados como o Darwin Core Archive (DwC-A), enormes volumes de dados brutos existem hoje na web. O Global Biodiversity Information Facility (GBIF) corresponde ao maior e mais diverso agregador, atualmente com aproximadamente 1,9 bilhões de registros de ocorrência de aproximadamente 1600 instituições.

PADRÕES DE DADOS E METADADOS DE BIODIVERSIDADE

A importância do uso de padrões e protocolos que sejam adequados ao contexto e sigam semântica correspondente, se baseia na redução de risco de erros, na interoperabilidade, escalabilidade de sistemas e consumo de dados por terceiros. No contexto de biodiversidade, os padrões de dados são primariamente definidos pela organização Biodiversity Information Standards (TDWG).

Neste tópico serão descritos, de maneira sucinta, os principais padrões de dados e metadados utilizados na gestão de dados e informação sobre biodiversidade. Optou-se por não se aprofundar neste tema e indicar o livro desenvolvido no âmbito do Ministério do Meio Ambiente, Diretrizes para a Integração de Dados de Biodiversidade (SILVA et al., 2015), que possui um excelente nível de detalhes e descrições detalhadas sobre o assunto.

No escopo deste trabalho, vamos considerar os dois padrões de dados, Darwin Core Archive (DwC-A) (WIECZOREK et al., 2012) e Dublin Core (WEIBEL;

---

https://www.gbif.org/

KOCH, 2000), e os padrões de metadados Ecological Metadata Language (EML) (VITOUSEK et al., 2005) e padrão INDE para dados e metadados geoespaciais .

Objetivamente, o DwC-A é o padrão utilizado para padronização de dados como registros de ocorrências e listas taxonômicas. É composto por um arquivo central, ou núcleo do recurso, onde as informações básicas do táxon ou da ocorrência são descritas, e arquivos chamados extensões que abrigam informações complementares. Este conjunto de arquivos se relacionam pelo identificador único de cada táxon ou ocorrência, presente no arquivo central (Figura 2). O padrão de metadados EML, no escopo deste diagnóstico, é utilizado para descrever os "dados sobre os dados" que foram padronizados em DwC-A, ou seja, funciona para descrever os conjuntos de dados de ocorrência e listas taxonômicas, organizando as informações relacionadas a estes, funcionando como um documento anexo aos recursos de informação.

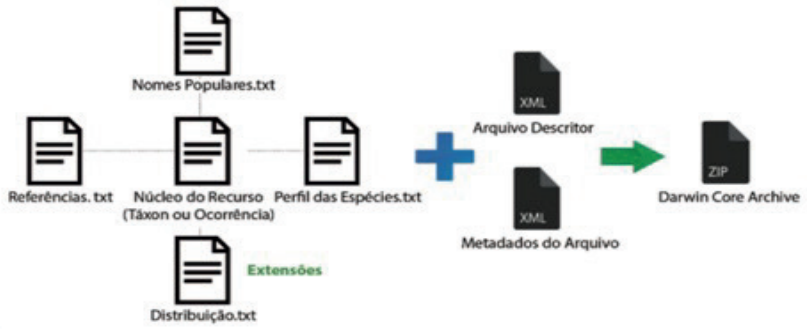

Figura 2. Componentes de um arquivo Darwin Core Archive (DwC-A). Retirado de Silva et al. (2015).

Integrados por padrões de dados como o Darwin Core Archive (DwC-A), enormes volumes de dados brutos existem hoje na web. O Global Biodiversity Information Facility (GBIF) corresponde ao maior e mais diverso agregador, atualmente com aproximadamente 1,9 bilhões de registros de ocorrência de aproximadamente 1600 instituições.

O padrão INDE para dados e metadados geoespaciais visa descrever as principais características e limitações dos dados geoespaciais, bem como ser uma documentação consistente que possibilite seu uso correto por parte da comunidade de usuários.

---

https://www.inde.gov.br/CatalogoMetadados

PRINCÍPIOS FAIR

Os princípios FAIR foram elaborados em 2014 por um grupo heterogêneo de pessoas do setor acadêmico e da iniciativa privada em um workshop em Leiden, na Holanda, para tratar dos obstáculos na busca e reutilização de dados, não somente de biodiversidade. Este grupo deliberou que o descobrimento, acesso, integração e reúso apropriado, assim como uma correta citação dos dados deveria ser baseada em uma série de acordos comunitários de princípios e práticas vastamente aceitos (WILKINSON et al., 2016).

A iniciativa está sendo implementada globalmente por meio do GO FAIR . No Brasil, o escritório GO FAIR Brazil está implementando diversas frentes e aguarda-se o lançamento de uma frente para biodiversidade.

Os princípios FAIR resumem muito bem as discussões sobre gestão e compartilhamento de dados que vêm sendo feitas nas últimas décadas e que direcionam os princípios da Informática da Biodiverside (BISBY, 2000; GALHAet l., 2020; SOBERÓN; PETERSON, 2004). Estes princípios indicam que os dados devem ser, em inglês, Findable (Buscáveis), Accessible (Acessíveis), Interoperable (Interoperáveis) e Reusable (Reutilizáveis), tanto para pessoas quanto para máquinas. A Figura 3 ilustra os principais pontos relacionados à gestão de dados seguindo os princípios FAIR.

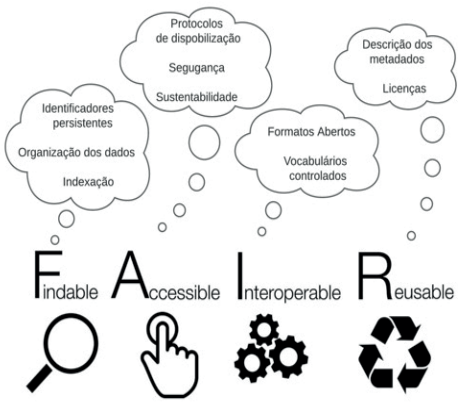

Figura 3. Princípios FAIR de compartilhamento de dados

PUBLICAÇÃO DE DADOS PARA PESSOAS E MÁQUINAS

A disponibilização ou publicação de dados acompanhados de metadados completos é o ponto principal quando se trata do uso da informação, seja pela ciência, por tomadores de decisão ou pela sociedade em geral. Ela constitui também o princípio básico da informática da biodiversidade e sua realização de forma adequada é o pilar fundamental dos princípios FAIR de compartilhamento de dados, que inclusive aponta que estes devem ser acessíveis para máquinas e humanos.

https://www.gbif.org/

Dados para humanos via repositórios de dados

A publicação de dados para humanos implica em que estes devam estar acessíveis e em formatos editáveis. Recomenda-se que sejam publicados via repositórios genéricos ou temáticos, para categorias específicas de dados.

A Figura 4, adaptada de (DALCIN et al., 2019), indica o conjunto de repositórios de dados utilizados pelo Jardim Botânico do Rio de Janeiro (JBRJ), para publicação "para humanos" por categoria (dados estruturados, não estruturados, multimídia e geográficos). Apesar de este ser um estudo de caso realizado para atender as demandas internas de uma instituição, e de que as tecnologias avançam e se substituem com frequência, este modelo pode servir para que gestores compreendam as aplicações do uso de repositórios de dados.

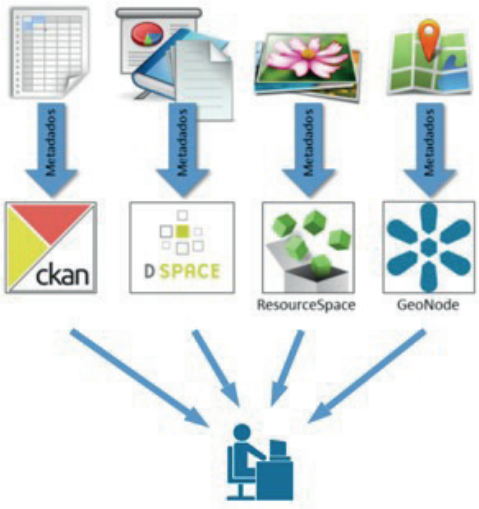

Figura 4. Publicação de dados via repositórios heterogêneos distribuídos no JBRJ (adaptado de Dalcin et al., 2019)

Dados de biodiversidade via Integrated Publishing Toolkit (IPT)

Em se tratando de dados taxonômicos e de ocorrência biológica, é fundamental que sigam padrões adequados e sejam publicados via repositório temático específico para este tipo de dado. Atualmente o repositório mais indicado para publicação deste tipo é o Integrated Publishing Toolkit (IPT), uma ferramenta desenvolvida e mantida pelo GBIF que, por ofertar dados seguindo padrão conhecido, podem ter seus recursos de informação consumidos por máquinas, mas também por humanos. Esta ferramenta publica os dados estritamente no formato DwC-A