



XXI ENANCIB

Encontro Nacional de Pesquisa em Ciência da Informação

50 anos de Ciência da Informação no Brasil:
diversidade, saberes e transformação social

Rio de Janeiro • 25 a 29 de outubro de 2021

XXI Encontro Nacional de Pesquisa em Ciência da Informação – XXI ENANCIB

GT 2 – Organização e Representação do Conhecimento

INTEROPERABILIDADE SEMÂNTICA NO CONTEXTO DE DADOS DA BIODIVERSIDADE: UM ESTUDO DE CASO SOBRE A UTILIZAÇÃO DE PADRÕES DE METADADOS

SEMANTIC INTEROPERABILITY IN THE CONTEXT OF BIODIVERSITY DATA: A CASE STUDY ON THE USE OF METADATA STANDARDS

Filipi Miranda Soares - Universidade de São Paulo (USP)

Raíssa Yuri Hamanaka - Universidade Estadual de Londrina (UEL)

Benildes Coura Moreira dos Santos Maculan - Universidade Federal de Minas Gerais (UFMG)

Modalidade: Trabalho Completo

Resumo: Trata acerca da interoperabilidade entre distintos sistemas de informação sobre a biodiversidade, no que se refere à descrição de registros de ocorrência de espécies, tendo em vista o compartilhamento de dados. O objetivo do estudo foi analisar como ocorre a interoperabilidade entre a) um registro de ocorrência de espécie entre os ambientes informacionais iNaturalist (iNat) e Global Biodiversity Information Facility (GBIF) e b) os dados de interações biológicas entre o iNat e Global Biotic Interactions (GloBI). Os procedimentos metodológicos são de objetivos exploratórios e descritivos, com abordagem qualitativa, e adotou a técnica do estudo de caso único, analisando como ocorre a interoperabilidade semântica entre os ambientes informacionais selecionados. Foram realizadas análises dos dados importados pelo GBIF e pelo GloBI, tendo como origem os dados do iNat, destacando como a interoperabilidade ocorreu entre esses sistemas. Os resultados evidenciaram que o uso de padrões de metadados permite que diferentes ambientes informacionais compartilhem dados entre si, garantindo a interoperabilidade semântica entre seus metadados. Especificamente entre iNat-GBIF e iNat-GloBI, a interoperabilidade semântica ocorre por meio do uso de padrões de metadados e de vocabulários controlados. O sistema do GBIF reconheceu e indexou, automaticamente, o registro de dados criado no iNat com metadados Darwin Core (DwC). O mesmo aconteceu com o GloBI, que reconheceu e indexou o registro de dados do iNat, automaticamente. Como limitação, indica-se que a análise dos metadados se restringiu a um pequeno grupo, levando à sugestão de estudo futuro em um escopo mais abrangente, que envolva também outros padrões de metadados.

Palavras-Chave: interoperabilidade; padrão de metadados; biodiversidade.

Abstract: This study deals with the interoperability between different biodiversity information systems, with respect to the description of species occurrence records, with a view to sharing data. The objective of the study was to analyze how the interoperability occurs between informational environments: a) a species occurrence record between iNaturalist (iNat) and Global Biodiversity Information Facility (GBIF) and b) biological interactions data between iNaturalist (iNat) and Global Biotic Interactions (GloBI). The methodological procedures are of exploratory and descriptive objectives, with a qualitative approach, and adopted the single case study technique, analyzing how the semantic interoperability occurs between the selected informational environments. Analyses of

the data imported by GBIF and GloBI were carried out, having as origin the data from iNat, highlighting how interoperability occurred between these systems. The results showed that the use of metadata standards allows different information environments to share data among themselves, ensuring semantic interoperability between their metadata. Specifically, between iNat-GBIF and iNat-GloBI, semantic interoperability occurs through the use of metadata standards and controlled vocabularies. The GBIF system automatically recognized and indexed the data record created in iNat with Darwin Core (DwC) metadata. The same was true for GloBI, which automatically recognized and indexed the iNat data record. As a limitation, it is indicated that the metadata analysis was restricted to a small group, leading to the suggestion of future study in a broader scope, also involving other metadata standards.

Keywords: interoperability; metadata standard; biodiversity.

1 INTRODUÇÃO

A atual era da informação é caracterizada pela multiplicidade de dispositivos conectados em rede e gerando dados em diversos formatos, fenômeno conhecido como *big data*. A estruturação e mineração de grandes volumes de dados (*big data*) apresenta-se como um grande desafio, no qual enormes volumes de dados heterogêneos devem ser processados e acessados em alta velocidade, mantendo-se a confiabilidade dos dados, para posterior agregação de valor às informações prévias (RIBEIRO, 2014; WAMBA *et al.*, 2015). Nesse contexto de dados não-estruturados, tem origem um novo paradigma da ciência, orientado a dados, ou e-Science, paradigma este norteado pela necessidade de armazenamento, compartilhamento e reuso de dados científicos (SAYÃO; SALES, 2012).

Percebe-se que as tecnologias já são capazes de analisar volumes grandes de dados, porém, se eles são dados não-estruturados, ou seja, não possuem padrões para que a máquina consiga compreender e interpretar os dados, tal complexidade poderá resultar em perda desse conhecimento. Alves e Souza (2007) apontam que o uso de padrões de metadados, normas e modelos informacionais permitem a integração e o intercâmbio de dados (e informações) entre sistemas. Assim, é importante disponibilizar dados já estruturados, que sigam padrões predeterminados, que permitam a sua recuperação e interpretação, trazendo análises confiáveis. Sousa (2012, p. 68) destaca que uma recuperação satisfatória exige “[...] uma representação adequada” das informações que serão disponibilizadas nos sistemas de informação. Essa representação adequada de um dado pode ser entendida como uma “[...] representação convencional, codificada, de uma informação em forma que permita submetê-la a processamento eletrônico” (LE COADIC, 2004, p. 8).

No âmbito da Ciência da Informação, considera-se que a representação da informação pode ser realizada por elementos temáticos e descritivos. De acordo com Maimone, Silveira e Tálamo (2011), os dois tipos de representação ocorrem de maneira fluida para facilitar a recuperação da informação pelo usuário. Nesse sentido, é possível pensar em uma representação com maior valor semântico sobre a informação, que objetiva aumentar a precisão na descrição de recursos informacionais, e, conseqüentemente, reduzir a revocação de informações pelos usuários (PELLEGRINO *et al.*, 2017; RAMALHO; VIDOTTI; FUJITA, 2007). No contexto web, esse maior valor semântico pode ser feito em uma representação formal dos recursos informacionais, facilitando que máquinas possam captar informações temáticas e descritivas desse recurso.

Ribeiro (2015) identifica dois tipos de formalismos na representação de recursos informacionais: o nível verbal da sintaxe (nível lógico textual) e o nível de abstração (campo das ideias), que contribuem para que haja a interoperabilidade entre diferentes recursos (RAMALHO; VIDOTTI; FUJITA, 2007). Interoperabilidade pode ser definida como “[...] a capacidade de sistemas computacionais em trocar dados sem intervenção humana” (ALMEIDA, 2020, p. 27) ou, como “[...] a capacidade de sistemas autônomos (informatizados ou não) comunicarem de modo transparente entre si, devido à adoção de padrões comuns e protocolos que permitem o uso compartilhado de informações” (MOURA, 2009, p. 61). Almeida (2020) aponta que a falta de interoperabilidade entre sistemas de informação é um problema mundial, que não é recente, sendo que Sayão e Marcondes (2008) identificam seis tipos de interoperabilidade: 1) técnica (como o protocolo Z3950, ISO-ILL e XML); 2) semântica; 3) política/humana; 4) intercomunitária; 5) legal e 6) internacional.

Discorrendo sobre essa questão, Chan e Zeng (2006, p. 4-5, tradução nossa) indicam que a interoperabilidade ocorre em três diferentes níveis:

Nível de esquema: os esforços são focados nos elementos dos esquemas de metadados, sendo independente de qualquer aplicação. Os resultados geralmente aparecem como conjuntos de elementos derivados ou esquemas codificados, mapeamentos, perfis de aplicação e registros de elementos.

Nível de registro: os esforços visam integrar os registros de metadados através do mapeamento dos elementos de acordo com os significados semânticos desses elementos. Os resultados comuns incluem registros convertidos e novos registros resultantes da combinação de valores dos registros existentes.

Nível de repositório: com registros minerados ou integrados de fontes variadas, os esforços neste nível se concentram no mapeamento de cadeias de valor associadas a elementos específicos (por exemplo, termos associados

a elementos de assunto ou de formato). Os resultados permitem a pesquisa cruzada de coleções de dados.

Considera-se que o uso de padrão de metadados ocorre no nível verbal da sintaxe, de maneira formal e explícita, estando a interoperabilidade semântica, foco deste estudo, caracterizada pelo uso de padrões de metadados, classificações, tesouros e ontologias, na distinção dos significados dos recursos informacionais.

O nível do conteúdo dos recursos informacionais se norteia pelos dados e metadados, e envolvem acordos semânticos que vão orientar a interpretação das informações disponibilizadas. Para Fusco (2011, p. 54), o nível do conteúdo “remete à interoperabilidade semântica, na qual a representação e organização do conhecimento são áreas-chave a serem estudadas”. A interoperabilidade semântica no nível do esquema de metadados pode ter como função criar mapeamentos (*crosswalks*) entre diferentes vocabulários para que haja interoperabilidade entre eles. Na Figura 1, observa-se um exemplo de como esse mapeamento entre vocabulários pode ser feito.

Figura 1 – Exemplo de mapeamento entre vocabulários.

target source	Absolute crosswalking	Relative crosswalking
VAR Core (3.0)	Dublin Core	Dublin Core
<i>Technique</i>	-----	<i>Format</i>
<i>Location.Current Repository</i>	-----	<i>Contributor Coverage</i>

Fonte: Chan e Zeng (2006).

Basicamente, um mapeamento é um esforço de se definir a equivalência entre termos de dois esquemas de metadados diferentes para que repositórios aplicando esquemas de metadados distintos possam transmitir dados entre si. No exemplo da Figura 1, observa-se que os autores classificam o mapeamento entre os termos do esquema de metadados VAR Core e Dublin Core como absoluta ou parcial. De acordo com Chan e Zeng (2006), dificilmente o mapeamento entre dois elementos de esquemas diferentes será considerado absoluto, visto que o contexto de criação dos elementos de metadados são distintos, o que afeta diretamente a semântica dos elementos. Desta maneira, considera-se que os três níveis de interoperabilidade definidos por Chan e Zeng (2006) estão intrinsecamente relacionados: a interoperabilidade entre esquemas de metadados permite a criação de registros de metadados que compartilham uma semântica única entre si que, conseqüentemente, podem ser transmitidos entre diferentes repositórios de dados.

No contexto da biodiversidade, escolhido para a realização deste estudo de caso, esforços têm sido empregados em tornar os dados mais semanticamente estruturados de forma a permitir a interoperabilidade, como, por exemplo, o portal da Global Biodiversity Information Facility¹ (GBIF), o iNaturalist² (iNat) e a Global Biotic Interactions³ (GloBI).

Para que essas bases de dados sejam capazes de indexar e compartilhar dados, os padrões de metadados apresentam-se como uma das peças-chave na infraestrutura das bases. Dentre os estudos correlatos, que tratam de padrões de metadados no contexto da biodiversidade, destacam-se: o padrão Access to Biological Collection Data (ABCD⁴) (HOLETSCHEK *et al.*, 2012), o Darwin Core (DwC) (WIECZOREK *et al.*, 2012), o Ecological Metadata Language (EML) (FEGRAUS *et al.*, 2005) e o Genome Metadata aplicados no sistema de informação do PathoSystems Resource Integration Center (PATRIC) (DAVIS *et al.*, 2020). Padrões generalistas, como o Dublin Core (WEIBEL *et al.*, 1998), também podem ser aplicados na descrição de registros de dados de biodiversidade como campos de metadados complementares.

A partir da problematização exposta, que destaca a existência de elementos computacionais que possibilitam manter a interoperabilidade entre diferentes sistemas de informação, e, de maneira mais específica, nas questões da ciência da biodiversidade, que também lida com os problemas de interoperabilidade semântica, a presente pesquisa pretende responder: como o uso de padrões de metadados tem contribuído para a interoperabilidade entre sistemas de informação sobre biodiversidade? Para tanto, o objetivo da pesquisa é investigar como ocorre a interoperabilidade de um registro de ocorrência de espécie entre os ambientes informacionais iNat e GBIF e os dados de interações biológicas entre o iNat e GloBI.

2 MATERIAIS E MÉTODOS

Gil (1994) afirma que uma pesquisa pode ser caracterizada de acordo com seus objetivos, abordagem, natureza e técnicas. Seguindo essa premissa, esta pesquisa se qualifica como de objetivos exploratórios e descritivos, com abordagem qualitativa, natureza aplicada e utilizou a técnica do estudo de caso, pois tem por base a análise de um único registro de ocorrência de espécie.

¹ Disponível em: <https://www.gbif.org>. Acesso em: 21 maio 2021.

² Disponível em: <https://www.inaturalist.org>. Acesso em: 21 maio 2021.

³ Disponível em: <https://www.globalbioticinteractions.org>. Acesso em: 21 maio 2021.

⁴ Disponível em: <https://doi.org/10.1080/11263504.2012.740085>

O estudo de caso pode ser definido como “[...] um estudo intensivo de um caso singular ou de um pequeno número de casos que se baseia em dados e promessas de elucidar uma população maior de casos” (GERRING, 2019, p. 69). Segundo Yin (2015, p. 18-19), um estudo de caso pode ser delimitado conforme seu escopo e características: “[...] investiga um fenômeno contemporâneo (o ‘caso’) em profundidade e em seu contexto de mundo real”, estando os limites entre o fenômeno e o contexto não claramente evidentes, contando “[...] com múltiplas fontes de evidência”. A partir disso, este estudo de caso tem como escopo a descrição por metadados de um registro de ocorrência de espécie, e se caracteriza como uma investigação de um fenômeno contemporâneo, em profundidade e no contexto de mundo real.

2.1 Procedimentos do estudo de caso

O estudo de caso foi organizado em cinco etapas: 1) caracterização dos ambientes informacionais; 2) observação de uma espécie; 3) registro da espécie observada; 4) validação da espécie observada; 5) adequação do registro, descritas a seguir.

Etapla 1: caracterização dos ambientes informacionais: descrever os ambientes informacionais: a) plataforma iNat, b) o portal de dados do GBIF e c) a base de dados GloBI.

Etapla 2: observação de uma espécie: observação de uma borboleta da espécie *Callicore sorana*⁵ (vulgarmente conhecida como borboleta-oitenta), em ambiente natural, se alimentando de um romã, fruto de um indivíduo da espécie *Punica granatum*⁶, no município de Moeda, Minas Gerais; foi utilizada uma câmera fotográfica digital modelo SM-A715F para registro fotográfico da observação; o GPS do próprio equipamento foi utilizado para determinar a localização precisa (com coordenadas geográficas exatas) da observação; a observação foi feita no dia 27 de março de 2021, às 14h27min, horário de Brasília (GMT-3).

Etapla 3: registro da espécie observada: o registro da observação foi feito na plataforma iNaturalist.org, por meio de aplicativo disponível para Android; as quatro fotografias feitas durante a observação foram inseridas no aplicativo, que detectou, automaticamente, os dados de localização e a data da observação; o registro no iNat foi feito no dia 28 de março de 2021, às 16h39min, horário de Brasília (GMT-3).

⁵ *Callicore sorana* in GBIF Secretariat (2021). GBIF Backbone Taxonomy. Checklist dataset <https://doi.org/10.15468/39omei> accessed via GBIF.org on 2021-05-18.

⁶ Identificador do táxon: <https://www.inaturalist.org/taxa/58300-Punica-granatum>.

Etapa 4: validação da espécie observada: no dia 29 de março de 2021, um usuário especialista⁷ validou a identificação da espécie observada como *Callicore sorana*; a partir da validação da identificação, o registro se tornou elegível para ser importado pela base de dados GBIF, já que atendeu a todos os critérios de qualidade de dados do iNat⁸; de acordo com Ueda (2021, online), “Observações do iNat tornam-se candidatas para Grau de Pesquisa quando elas têm uma foto, data e coordenadas. Elas se tornam Grau de Pesquisa quando a comunidade concorda com a identificação da espécie”. Posteriormente, o registro de dados foi importado, automaticamente, pelo GBIF.

Etapa 5: adequação do registro: para que o registro se tornasse elegível para ser importado pela base de dados GloBI, foram acrescentados os pares de metadados e valores *Feeding on: Punica granatum*⁹ e *Interaction->Ate fruit of: Punica granatum*¹⁰ para descrever a interação da borboleta com a romãzeira no registro de dados, já que a base de dados GloBI rastreia apenas registros de dados que representam interações entre espécies. Os metadados adicionais foram inseridos na seção Campos de Observação (vide Figura 2), e o registro de dados foi importado, automaticamente, pelo GloBI.

3 RESULTADOS E DISCUSSÕES

A partir dos cinco procedimentos executados, esta seção apresenta os resultados e as respectivas discussões, que foram organizados em três subseções: a primeira descreve a caracterização dos ambientes informacionais, a segunda é dedicada à análise da interoperabilidade entre o iNat e o GBIF, e, a terceira, destaca a análise entre o iNat e o GloBI.

3.1 Caracterização dos ambientes informacionais

Aqui estão descritos os três ambientes informacionais utilizados no estudo: a) plataforma iNaturalist (iNat), b) o portal de dados do Global Biodiversity Information Facility (GBIF) e c) a base de dados Global Biotic Interactions (GloBI).

a) plataforma iNaturalist (iNat).

⁷ Perfil disponível em: <https://www.inaturalist.org/people/1296774>. Acesso em: 18 maio 2021.

⁸ Critérios de qualidade de dados: <https://www.inaturalist.org/pages/help#quality>. Acesso em: 18 maio 2021.

⁹ Definição: “Which species is the observed species feeding on?”. URI: https://www.inaturalist.org/observation_fields/1685.

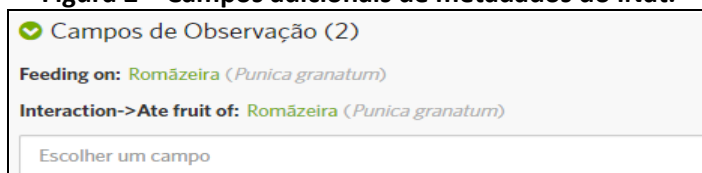
¹⁰ Definição: “The observed species was seen eating fruit of this plant”. URI: https://www.inaturalist.org/observation_fields/3128

O iNat é uma plataforma que obtém dados a partir da ciência cidadã, ou seja, do esforço de observação de espécies na natureza por pessoas sem formação científica.

Incentiva a participação de uma grande variedade de entusiastas da natureza [...]. Através da conexão dessas diferentes percepções e expertise do mundo natural, o iNaturalist espera criar uma ampla conscientização comunitária sobre a biodiversidade local e promover uma maior exploração dos ambientes locais (INATURALIST, 2021, tradução nossa).

O iNat possui, atualmente, quase dois milhões de usuários registrados como observadores, e mais de 77 milhões de registros de observações de espécies¹¹. Ele utiliza diversos vocabulários controlados (metadados e ontologias) para descrever os registros de dados. A versão Web do iNat permite ainda acrescentar metadados adicionais, além dos que são apresentados por *default* para criação dos registros (podem ser criados novos metadados ou utilizar metadados criados por outros usuários na plataforma), conforme a Figura 2.

Figura 2 – Campos adicionais de metadados do iNat.



Fonte: dados da pesquisa (2021).

b) portal de dados do Global Biodiversity Information Facility (GBIF).

O GBIF “[...] é uma rede internacional e infraestrutura de dados financiada pelos governos mundiais e destinada a fornecer a qualquer pessoa, em qualquer lugar, acesso aberto a dados sobre todas as formas de vida na Terra” (GLOBAL BIODIVERSITY INFORMATION FACILITY, 2021, tradução nossa). O GBIF oferece acesso a dados de ocorrência de espécies que, resumidamente, são dados da existência de determinada espécie em determinado local e período. Atualmente o portal GBIF.org possui mais de 1,8 bilhões de registros de dados¹².

c) base de dados Global Biotic Interactions (GloBI).

O GloBI é uma base de dados com escopo diferente do GBIF e iNat, apesar de também ser uma base de dados de biodiversidade. Enquanto o foco do iNat e do GBIF são dados espaço-temporais de ocorrência de espécies, o GloBI indexa apenas dados de interações biológicas entre espécies, que não precisam obrigatoriamente conter dados de localização ou data de

¹¹ Dados baseados em consulta feita ao portal iNaturalist.org em 18 de agosto de 2021.

¹² Dados baseados em consulta feita ao portal GBIF.org em 18 de agosto de 2021.

observação (POELEN; SIMONS; MUNGALL, 2014). O GloBI “[...] fornece acesso aberto a encontrar dados de interação de espécies (por exemplo, predador-presa, polinizador-planta, patógeno-hospedeiro, parasita-hospedeiro) combinando conjuntos de dados abertos existentes usando software de código aberto” (GLOBAL BIOTIC INTERACTIONS, 2021b, tradução nossa).

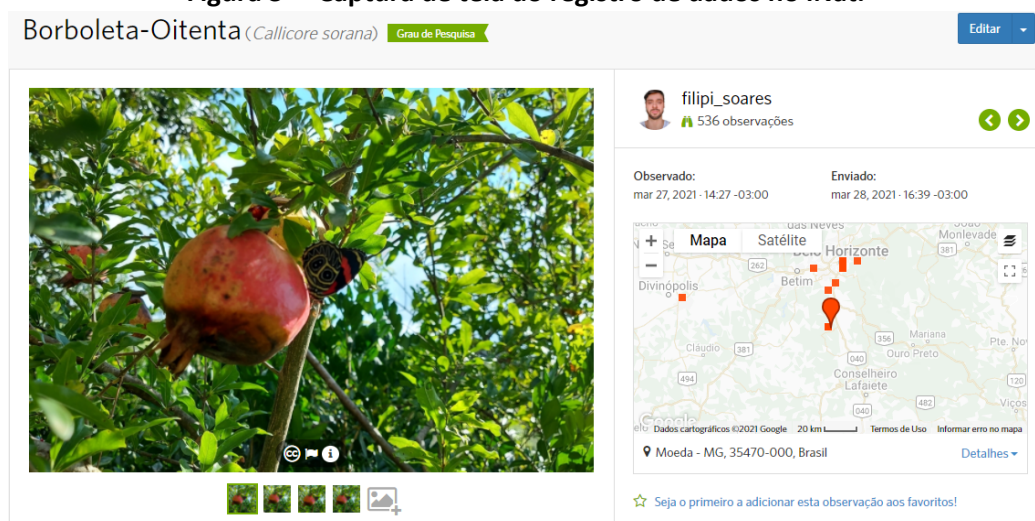
3.2 Interoperabilidade entre iNat e GBIF

Inicialmente, dois registros de dados distintos foram criados, com as seguintes URIs:

- a) iNat: <<https://www.inaturalist.org/observations/72295928>>;
- b) GBIF: <<https://www.gbif.org/occurrence/3070549701>>.

A Figura 3 é uma captura de tela de uma parte do registro de dados criado no iNat. Os demais detalhes podem ser visualizados ao se acessar o registro completo por meio da URI.

Figura 3 – Captura de tela do registro de dados no iNat.



Fonte: dados da pesquisa (2021).

O registro de metadados completo em formato CSV, com todos os metadados e valores utilizados para os fins deste estudo de caso, foi baixado do portal iNaturalist.org e pode ser acessado no material suplementar deste trabalho, vide Tabela 1S (SOARES; HAMANAKA; MACULAN, 2021). O registro de dados também pode ser visualizado diretamente na plataforma do iNat por meio da URI informada anteriormente nesta seção.

O registro de dados foi importado pelo GBIF, a partir da própria infraestrutura de *crawling* da base¹³, que rastreia periodicamente os registros de dados com Grau de Pesquisa no iNat. Para facilitar a citação dos dados, o GBIF permite que um *dataset* gerado a partir de uma

¹³ GLOBAL BIODIVERSITY INFORMATION FACILITY. GBIF data processing. Copenhagen. Disponível em: <https://www.gbif.org/pt/data-processing>. Acesso em: 18 maio 2021.

query receba um DOI. Dessa forma, foi executada uma *query* na busca avançada no portal do GBIF, para localizar o registro de dados criado após a importação, com os seguintes critérios: (Scientific name: Callicore sorana AND Recorded by: Filipi Miranda Soares AND Event date: 2021-03-27). O conjunto de dados gerado a partir desta *query* está disponível no material suplementar deste trabalho (vide SOARES, 2021).

Como já abordado anteriormente, para que a interoperabilidade aconteça entre os repositórios, um dos elementos necessários é que compartilhem um mesmo conjunto de metadados ou, caso apliquem esquemas diferentes, que haja um mapeamento entre eles (CHAN; ZENG, 2006). A vista disso, foi possível observar como esse tipo de interoperabilidade ocorre entre o GBIF e o iNat. O Quadro 1S do material suplementar (SOARES; HAMANAKA; MACULAN, 2021) apresenta alguns exemplos de equivalências entre metadados do iNat e do GBIF.

Todos os metadados do GBIF, relacionados no Quadro 1S, apresentam o formato especificado pelo esquema de metadados DwC¹⁴, que é o responsável pela interoperabilidade dos metadados e dados entre os sistemas. A partir do Darwin Core Archive (DwCA¹⁵) — que é um arquivo padrão que organiza diferentes registros de metadados DwC com o auxílio do esquema EML (FEGRAUS *et al.*, 2005) — gerado pelo iNat, o sistema do GBIF reconheceu e importou, automaticamente, todos os dados da observação, conforme o escopo de dados do DwC (há metadados e dados específicos do iNat que não são importados pelo GBIF, já que dizem respeito a critérios de qualidade ou outros dados de controle que são internos do iNat).

Ainda sobre o Quadro 1S, foi possível observar que os nomes dos metadados no iNat apresentam formato diferente dos metadados do GBIF, e, conseqüentemente, do DwC. Os nomes dos metadados em DwC são formatados em *lowerCamelCase*, que é a escrita de palavras compostas, sem espaçamento entre os termos, utilizando apenas caracteres maiúsculos para identificar a inicial de um termo, exceto para o primeiro, que começa em minúsculo. Entretanto,

¹⁴ Disponível em: <https://dwc.tdwg.org/terms/>. Acesso em: 21 maio 2021.

¹⁵ “Darwin Core Archive (DwC-A) é um padrão de dados de informática da biodiversidade que faz uso dos termos do Darwin Core para produzir um único conjunto de dados para compartilhamento de dados de nível de espécies (taxonômicos), dados de ocorrência de espécies e dados de eventos amostrais. Um arquivo é um conjunto de arquivos de texto, em formato CSV, com um arquivo descritor (denominado meta.xml) para informar os outros como os arquivos são organizados” (REMSEN *et al.*, 2021, tradução nossa).

no momento de exportar os dados para o GBIF, o sistema do iNat faz o mapeamento dos metadados e gera um DwCA. Isso permite que o GBIF reconheça os metadados.

Foi possível observar, também, que em alguns casos o GBIF fez inferências ou modificou dados importados. No Quadro 1S, observa-se que o metadado *observed_on* do iNat não possui equivalente direto no GBIF. Entretanto, o GBIF apresenta a data de observação subdividida em três campos de metadados (*day*, *month* e *year*), que, possivelmente, foram inferidos a partir do campo *observed_on*, ou de algum outro campo do iNat, que também apresenta a data completa da observação. Para todos os campos de metadados, o GBIF exibe o valor interpretado (após a importação) e o valor original, conforme consta no registro do iNat e demonstrado no Quadro 1.

Quadro 1 – Dados de evento importados pelo GBIF.

Event			
Term	Interpreted	Original	Remarks
Day	27		Inferred
Month	3		Inferred
Year	2021		Inferred
Event date	2021-03-27T14:27:05	2021-03-27T14:27:05-03:00	Altered
Event time	17:27:05Z	17:27:05Z	
Verbatim event date	2021-03-27 2:27:05 PM GMT-03:00	2021-03-27 2:27:05 PM GMT-03:00	

Fonte: dados da pesquisa (2021).

No Quadro 1, observa-se que, na coluna *remarks*, o registro de metadados do GBIF informa qual o tipo de interpretação o dado sofreu, que pode ser uma inferência a partir de outros dados (*inferred*). Ou seja, o GBIF gerou um dado novo a partir dos dados originais do iNat, ou uma alteração (*altered*), que indica que o dado original de um campo específico de metadados foi modificado. Em outros casos, algum dado pode ser excluído ou até mesmo arredondado ou presumido, como as coordenadas geográficas (vide Quadro 2).

Quadro 2 – Dados de localização importados pelo GBIF.

Location			
Term	Interpreted	Original	Remarks
Country or area	Brazil		Inferred
Country code	BR	BR	
Decimal latitude	-20.290667	-20.2906666997	Coordinate rounded
Decimal longitude	-43.992761	-43.992761	Coordinate rounded
Geodetic datum	WGS84		Geodetic datum assumed WGS84
State province	Minas Gerais	Minas Gerais	
Verbatim locality	Moeda - MG, 35470-000, Brasil	Moeda - MG, 35470-000, Brasil	

Fonte: dados da pesquisa (2021).

Foi possível observar que os dados importados do iNat foram automaticamente reconhecidos pelo GBIF, sendo possível graças à aplicação de padrões de metadados, além de

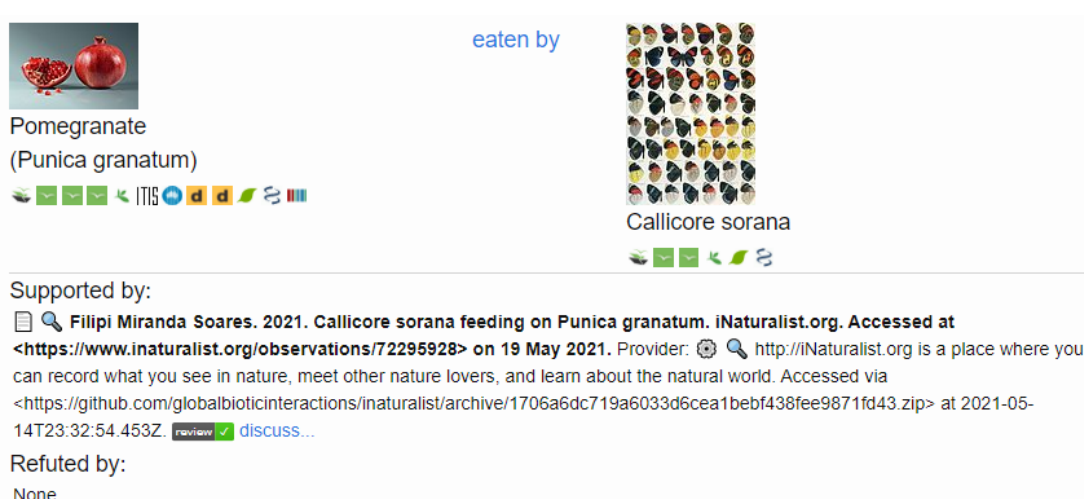
padrões para codificação de valores (como o formato de datas) e à infraestrutura de *crawling* da base.

3.3 Interoperabilidade entre iNat e GloBI

No caso do GloBI, os dados são indexados a partir de outros *datasets*, portanto, a URI do *dataset* original também é incorporada, para facilitar a recuperação da informação. Utilizando a URI <<https://www.inaturalist.org/observations/72295928>> no campo de busca *according to*, foi possível recuperar os dados da interação registrada neste estudo¹⁶. A Figura 4 mostra os dados indexados pelo GloBI, a partir do iNat.

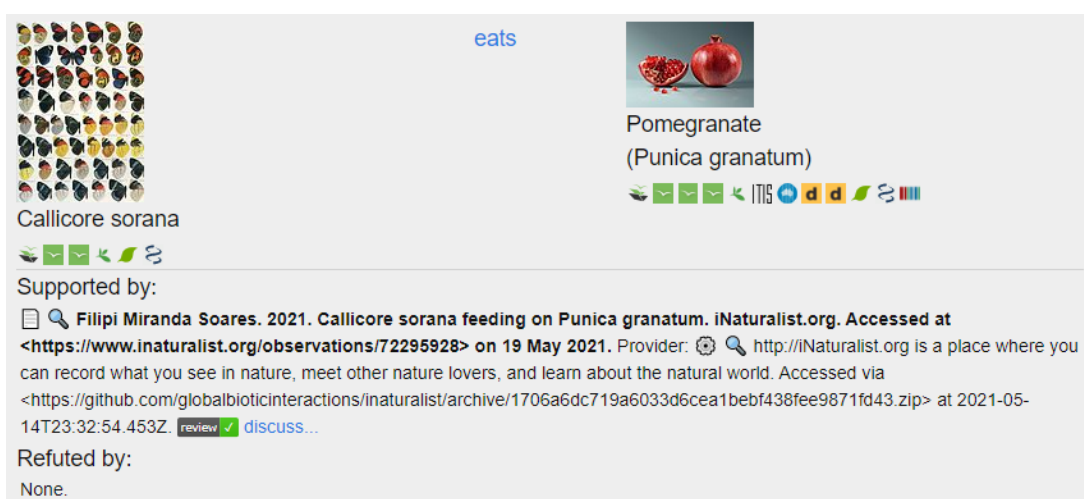
Figura 4 – Dados de interações importados pelo GloBI.

a)



The screenshot shows an iNaturalist observation titled "Pomegranate (Punica granatum)". The image shows a pomegranate. The text "eaten by" is displayed in blue. To the right, a grid of butterfly icons is shown, with "Callicore sorana" identified below it. The observation is supported by a citation: "Filipi Miranda Soares. 2021. Callicore sorana feeding on Punica granatum. iNaturalist.org. Accessed at <<https://www.inaturalist.org/observations/72295928>> on 19 May 2021. Provider: http://iNaturalist.org is a place where you can record what you see in nature, meet other nature lovers, and learn about the natural world. Accessed via <<https://github.com/globalbioticinteractions/inaturalist/archive/1706a6dc719a6033d6cea1bebf438fee9871fd43.zip>> at 2021-05-14T23:32:54.453Z. review discuss...". The observation is not refuted.

b)



The screenshot shows an iNaturalist observation titled "Callicore sorana". The image shows a grid of butterfly icons. The text "eats" is displayed in blue. To the right, a pomegranate is shown, with "Pomegranate (Punica granatum)" identified below it. The observation is supported by the same citation as in (a): "Filipi Miranda Soares. 2021. Callicore sorana feeding on Punica granatum. iNaturalist.org. Accessed at <<https://www.inaturalist.org/observations/72295928>> on 19 May 2021. Provider: http://iNaturalist.org is a place where you can record what you see in nature, meet other nature lovers, and learn about the natural world. Accessed via <<https://github.com/globalbioticinteractions/inaturalist/archive/1706a6dc719a6033d6cea1bebf438fee9871fd43.zip>> at 2021-05-14T23:32:54.453Z. review discuss...". The observation is not refuted.

Fonte: dados da pesquisa (2021).

¹⁶ Resultado da busca disponível em:

<https://www.globalbioticinteractions.org/?accordingTo=https%3A%2F%2Fwww.inaturalist.org%2Fobservations%2F72295928&interactionType=interactsWith>. Acesso em: 18 maio 2021.

Conforme mencionado na etapa 5 dos procedimentos deste estudo de caso (seção **Erro! Fonte de referência não encontrada.**), foram adicionados os pares de metadados e valores *Feeding on: Punica granatum* e *Interaction->Ate fruit of: Punica granatum* no registro de dados do iNat, para que ele fosse rastreado pelo GloBI, visando atingir a interoperabilidade em nível de registro (CHAN; ZENG, 2006). Entretanto, no registro da Figura 4 não aparecem esses metadados, apenas os termos *eaten by* (parte A da Figura 4) e *eats* (parte B da Figura 4), que não foram adicionados no registro do iNat (vide Figura 2, na seção 3.1). Isso acontece porque o GloBI possui um vocabulário controlado que mapeia alguns metadados inseridos nos Campos de Observação do iNat (em linguagem natural), que descrevem interações ecológicas. Esse vocabulário controlado está disponível em formato CSV e pode ser consultado diretamente no GitHub (GLOBAL BIOTIC INTERACTIONS, 2021a). O Quadro 3 mostra os termos equivalentes para os dois metadados utilizados no registro do iNat.

Quadro 3 – Equivalência de metadados iNat-GloBI.

iNat		GloBI	
Observation field name	Observation field ID	Interaction type label	Interaction type ID
Feeding on	https://www.inaturalist.org/observation_fields/1685	eats	http://purl.obolibrary.org/obo/RO_0002470
Interaction: Fruit eaten by	https://www.inaturalist.org/observation_fields/3127	eaten by	http://purl.obolibrary.org/obo/RO_0002471

Fonte: Global Biotic Interactions, (2021a).

A partir da observação do Quadro 3, fica evidente o mapeamento entre os metadados, conforme definição de Chan e Zeng (2006). O GloBI utiliza os termos de interações ecológicas da Relation Ontology¹⁷ (RO) para estabelecer os relacionamentos entre os registros de dados e, apesar de não definir o nível de equivalência entre os metadados do iNat e os termos da RO, conforme sugerido por Chan e Zeng (2006), vide Figura 1, é possível perceber certo grau de equivalência semântica entre os termos a partir das *labels* atribuídas a eles. Destaca-se que 174 metadados do iNat foram mapeados pelo GloBI e encontraram correspondência com 20 termos equivalentes da RO — sendo esses 20 termos aplicados pelo GloBI —, conforme lista no Quadro 2S do material suplementar (SOARES; HAMANAKA; MACULAN, 2021).

Foi possível constatar que também há interoperabilidade entre os metadados do iNat e do GloBI, graças ao mapeamento entre os vocabulários de ambas as bases. O GloBI adota, como parte de seu esquema de metadados, os termos de uma ontologia pré-existente, da RO, o que

¹⁷ Disponível em: <http://www.ontobee.org/ontology/RO>. Acesso em: 22 maio 2021.

representa um ganho operacional, visto que o reuso da ontologia poupou a equipe de desenvolvimento da base de desenvolver um novo vocabulário controlado.

4 CONSIDERAÇÕES FINAIS

Este estudo partiu da problematização sobre a existência de elementos computacionais que possibilitam manter a interoperabilidade entre diferentes sistemas de informação, em especial no âmbito da biodiversidade, analisando como o uso de padrões de metadados tem contribuído para a interoperabilidade entre três diferentes ambientes informacionais iNat e GBIF e GloBI.

Os resultados demonstraram que o uso de padrões de metadados permite que diferentes ambientes informacionais compartilhem dados entre si, e, conseqüentemente, que haja interoperabilidade entre seus metadados (CHAN; ZENG, 2006; CASTRO; SANTOS, 2014). A análise realizada evidenciou que a interoperabilidade semântica entre o iNat-GBIF e iNat-GloBI ocorre por meio do uso de padrões de metadados entre os dois ambientes informacionais, o que garantiu o estabelecimento de correspondência entre metadados, e, assim, o compartilhamento de significado (valor) dos dados entre esses ambientes.

Confirmando as afirmações de Alves e Souza (2007), que destacam o uso de metadados, normas e modelos informacionais na integração e na troca de dados entre sistemas, verificou-se que as três plataformas estudadas utilizam padrões de metadados em comum, o que permite a interoperabilidade semântica. Além disso, utilizam APIs que reconhecem esses metadados e fazem a extração automática dos dados de uma plataforma para outra. Também foi possível evidenciar que o uso de URIs e de ontologias na explicitação formal de relacionamentos no registro de ocorrência de espécie do iNat permitiu que este fosse rastreado e indexado pelo GloBI e pelo GBIF. A interoperabilidade semântica entre os metadados dos ambientes informacionais investigados facilita a consulta, em uma única vez, desse conjunto de bases de dados, de forma integrada e com esforço reduzido, desafio que a interoperabilidade pretende cumprir, segundo Marcondes e Sayão (2001).

A limitação da pesquisa foi a análise dos metadados ter se restringido a um pequeno grupo, recorte que foi feito como exemplificação. O padrão Dublin Core, por exemplo, é aplicado pelo iNat e pelo GBIF para descrever as fotografias do registro de dados, e não foi abordado neste trabalho, além de outros metadados relativos a *copyright*. Logo, sugere-se como estudo

futuro a replicação deste estudo em um escopo mais abrangente, que envolva outros padrões de metadados.

AGRADECIMENTOS

O primeiro autor agradece à Capes, FAPESP e ao USP-FAPESP-IBM Center for Artificial Intelligence (C4AI) por todo o apoio técnico e financeiro concedido ao desenvolvimento da pesquisa de doutorado do autor. A terceira autora agradece ao CNPq pelo apoio à pesquisa no Processo 303650/2019-2.

MATERIAL SUPLEMENTAR

SOARES, Filipi Miranda; HAMANAKA, Raíssa Yuri; MACULAN, Benildes Coura Moreira dos Santos. Dataset "Interoperabilidade semântica no contexto de dados da biodiversidade: um estudo de caso sobre a utilização de padrões de metadados". **Mendeley Data**, v. 1, Aug. 2021. DOI: <https://doi.org/10.17632/9dfmcmkdn7.1>.

SOARES, Filipi Miranda. **GBIF Occurrence Download**. GBIF.org, 21 May 2021. DOI: <https://doi.org/10.15468/dl.fkvake>.

REFERÊNCIAS

ALMEIDA, M. B. **Ontologia em Ciência da Informação**: teoria e método. Curitiba: CRV, 2020. 374 p. (Coleção Representação do Conhecimento em Ciência da Informação, v. 1). DOI: <https://doi.org/10.24824/978655578679.8>.

ALVES, M. D. R.; SOUZA, M. I. F. Estudo da correspondência de elementos metadados: Dublin Core e MARC 21. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 4, n. 2, p. 20-38, jan./jun. 2007.

CASTRO, F. F.; SANTOS, P. L. V. A. C. Elementos de interoperabilidade na perspectiva da catalogação descritiva. **Informação & Sociedade: Estudos**, João Pessoa, v. 24, n. 3, p. 13-25, set./dez. 2014.

CHAN, L.; ZENG, M. Metadata interoperability and Standardization: a study of methodology: part I. **D-Lib Magazine**, v. 12, n. 6, p. 1-19, June 2006. DOI: <https://doi.org/10.1045/june2006-chan>.

DAVIS, J. J. *et al.* The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. **Nucleic Acids Research**, v. 48, n. D1, p. D606-D612, Jan. 2020. DOI: <https://doi.org/10.1093/nar/gkz943>

FEIGRAUS, E. H. *et al.* 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**, v. 86, n. 3, p. 158-168, July 2005. DOI: [https://doi.org/10.1890/0012-9623\(2005\)86\[158:MTVOED\]2.0.CO;2](https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2)

FUSCO, E. **Aplicação dos FRBR na modelagem de catálogos bibliográficos digitais**. São Paulo: Cultura Acadêmica, 2011.

GERRING, J. **Pesquisa de estudo de caso**: princípios e práticas. Petrópolis, RJ: Vozes, 2019. 359 p.

- GIL, A. C. **Métodos e técnicas de pesquisa social**. 4. ed. São Paulo: Atlas, 1994.
- GLOBAL BIODIVERSITY INFORMATION FACILITY. **What is GBIF?** Disponível em: <https://www.gbif.org/what-is-gbif>. Acesso em: 21 maio 2021.
- GLOBAL BIOTIC INTERACTIONS. [inaturalist/interaction_types.csv](https://github.com/globalbioticinteractions/inaturalist/blob/main/interaction_types.csv). Disponível em: https://github.com/globalbioticinteractions/inaturalist/blob/main/interaction_types.csv. Acesso em: 21 maio 2021a.
- GLOBAL BIOTIC INTERACTIONS. **What is GloBI?** Disponível em: <https://www.globalbioticinteractions.org/about>. Acesso em: 21 maio 2021b.
- HOLETSCHEK, J. *et al.* The ABCD of primary biodiversity data access. **Plant Biosystems**, v. 146, n. 4, p. 771-779, 2012. DOI: <https://doi.org/10.1080/11263504.2012.740085>.
- INATURALIST. **What is iNaturalist?** Disponível em: <https://www.inaturalist.org/pages/help#general1>. Acesso em: 21 maio 2021.
- LE COADIC, Y. **A ciência da informação**. 2. ed. Brasília: Briquet de Lemos, 2004. 124 p.
- MAIMONE, G. D.; SILVEIRA, N. C.; TÁLAMO, M. F. G. M. Reflexões acerca das relações entre representação temática e descritiva. **Informação & Sociedade: Estudos**, João Pessoa, v. 21, n. 1, p. 27-35, 2011.
- MARCONDES, C. H.; SAYÃO, L. F. Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 24-33, set./dez. 2001.
- MOURA, M. A. Informação, ferramentas ontológicas e redes sociais *ad hoc*: a interoperabilidade na construção de tesouros e ontologias. **Informação & Sociedade: Estudos**, João Pessoa, v. 19, n. 1, p. 59-73, jan./abr. 2009.
- PELLEGRINO, A. L. *et al.* Bibliotecas e instituições de memória na web, dados ligados e web semântica: diálogos interdisciplinares. **Memória e Informação**, Rio de Janeiro, v. 1, n. 1, p. 53-72, 2017.
- POELEN, J. H.; SIMONS, J. D.; MUNGALL, C. J. Global biotic interactions: an open infrastructure to share and analyze species-interaction datasets. **Ecological Informatics**, v. 24, p. 148-159, Nov. 2014. DOI: <https://doi.org/10.1016/j.ecoinf.2014.08.005>.
- RAMALHO, R. A. S.; VIDOTTI, S. A. B. G.; FUJITA, M. S. L. Web semântica: uma investigação sob o olhar da Ciência da Informação. **DataGramaZero: Revista de Ciência da Informação**, v. 8, n. 6, 2007.
- REMSEN, D. *et al.* **Darwin Core Archives**: how-to guide. Version 2.1, released on 8 February 2021. Copenhagen: Global Biodiversity Information Facility, 2021. Disponível em: <https://github.com/gbif/ipt/wiki/DwCAHowToGuide>. Acesso em: 18 ago. 2021.
- RIBEIRO, C. J. S. Big Data: os novos desafios para o profissional da informação. **Informação & Tecnologia**, João Pessoa, v. 1, n. 1, p. 96-105, 2014.
- RIBEIRO, C. J. S. Uma investigação sobre o uso de vocabulários e formalismos: em busca de um caminho para a representação semântica de patrimônio digital e ativos de informação cultural. *In*: ENCONTRO DE EDUCAÇÃO E INVESTIGAÇÃO EM CIÊNCIA DA INFORMAÇÃO DA IBEROAMÉRICA E CARIBE, 7., 2015, Madrid. **Anais [...]** Madrid: Universidad Complutense de Madrid, 2015.

SAYÃO, L. F.; MARCONDES, C. H. O desafio da interoperabilidade e as novas perspectivas para as bibliotecas digitais. **TransInformação**, Campinas, v. 20, n. 2, p. 133-148, 2008. Disponível em: <https://goo.gl/z7AWxm>. Acesso em: 7 abr. 2021.

SAYÃO, L. F.; SALES, L. F. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação & Sociedade: Estudos**, João Pessoa, v. 22, n. 3, p. 179-191, 2012.

SOUSA, M. R. F. O acesso a informações e a contribuição da arquitetura da informação, usabilidade e acessibilidade. **Informação & Sociedade: Estudos**, João Pessoa, v. 22, n. esp., p. 65-76, 2012.

UEDA, K. **iNaturalist Research-grade Observations**. iNaturalist.org. 2021. Occurrence dataset <https://doi.org/10.15468/ab3s5x> accessed via GBIF.org on 18 May 2021.

WAMBA, S. F. *et al.* How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. **International Journal Production Economics**, v. 165, p. 234-246, 2015.

WEIBEL, S. *et al.* **Dublin Core Metadata for Resource Discovery**. Network Working Group, Sept. 1998. Disponível em: <https://www.hjp.at/doc/rfc/rfc2413.html>. Acesso em: 22 maio 2021.

WIECZOREK, J. *et al.* Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. **PLoS One**, v. 7, n. 1, Jan. 2012. DOI: <https://doi.org/10.1371/journal.pone.0029715>

YIN, R. K. **Estudo de caso: planejamento e métodos**. 5. ed. Porto Alegre: Bookman, 2015. 290 p.