DANIEL LINS DA SILVA

ESTRATÉGIA COMPUTACIONAL PARA APOIAR A REPRODUTIBILIDADE E
REUSO DE DADOS CIENTÍFICOS BASEADO EM METADADOS DE
PROVENIÊNCIA

São Paulo

2017

DANIEL LINS DA SILVA

ESTRATÉGIA COMPUTACIONAL PARA APOIAR A REPRODUTIBILIDADE E
REUSO DE DADOS CIENTÍFICOS BASEADO EM METADADOS DE
PROVENIÊNCIA

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
título de Doutor em Ciências.

São Paulo

2017

DANIEL LINS DA SILVA

ESTRATÉGIA COMPUTACIONAL PARA APOIAR A REPRODUTIBILIDADE E
REUSO DE DADOS CIENTÍFICOS BASEADO EM METADADOS DE
PROVENIÊNCIA

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
título de Doutor em Ciências.

Área de Concentração: Engenharia de
Computação

Orientador: Prof. Dr. Pedro Luiz Pizzigatti
Corrêa

São Paulo

2017

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 10 de Agosto de 2017

Assinatura do autor: Daniel Lins da Silva

Assinatura do orientador: Pedro L. Bonini

Catlogação-na-publicação

da Silva, Daniel Lins

ESTRATÉGIA COMPUTACIONAL PARA APOIAR A
REPRODUTIBILIDADE E REUSO DE DADOS CIENTÍFICOS BASEADO EM
METADADOS DE PROVENIÊNCIA / D. L. da Silva -- versão corr. -- São Paulo,
2017.

180 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo.
Departamento de Engenharia de Computação e Sistemas Digitais.

1.Proveniência dos Dados 2.Ciência dos Dados 3.Metadados
4.Pesquisas Reproduzíveis 5.Reuso de Dados I.Universidade de São Paulo.
Escola Politécnica. Departamento de Engenharia de Computação e Sistemas
Digitais II.t.

A Naiandra e ao Daniel, esposa e filho, pelo amor diário, apoio incondicional e pelo sacrifício das horas de convívio familiar. A Marilene e Teodorico, meus pais, pelos ensinamentos, pelas palavras de confiança e pelo suporte nos momentos que as dificuldades se mostravam mais fortes do que a vontade de continuar.

AGRADECIMENTOS

Primeiramente a Deus, por me permitir fazer parte deste contexto, no qual tenho a oportunidade de avançar na minha instrução e contribuir em soluções relevantes para a comunidade científica e a sociedade.

A minha família e meus amigos por todo o suporte, apoio e compreensão em todos os momentos. Tudo isso me deu força e motivação para enfrentar os desafios durante este longo período de doutoramento.

Ao Prof. Dr. Pedro Luiz Pizzigatti Corrêa, pelos ensinamentos, apoio, paciência e incentivo durante todo o período desta jornada, no qual nos tornamos verdadeiros amigos.

Aos professores Prof. Dr. Antônio Mauro Saraiva, Prof. Dr. Marcelo Morandini e Prof. Dra. Liria Matsumoto Sato pelo apoio na realização da pesquisa.

Aos parceiros de pesquisa e amigos do Laboratório de Automação Agrícola (LAA), do Núcleo de Pesquisa em Biodiversidade e Computação (Biocomp) e do Grupo de Estudos, Pesquisa e Extensão em Big Data.

À FAPEAM – Fundação de Amparo à Pesquisa do Estado do Amazonas – pela concessão da bolsa de doutorado, Edital N. 005/2012.

"A mente que se abre a uma nova ideia,
jamais voltará ao seu tamanho original."

Albert Einstein

RESUMO

A ciência moderna, apoiada pela *e-science*, tem enfrentado desafios de lidar com o grande volume e variedade de dados, gerados principalmente pelos avanços tecnológicos nos processos de coleta e processamento dos dados científicos. Como consequência, houve também um aumento na complexidade dos processos de análise e experimentação. Estes processos atualmente envolvem múltiplas fontes de dados e diversas atividades realizadas por grupos de pesquisadores geograficamente distribuídos, que devem ser compreendidas, reutilizadas e reproduzíveis. No entanto, as iniciativas da comunidade científica que buscam disponibilizar ferramentas e conscientizar os pesquisadores a compartilharem seus dados e códigos-fonte, juntamente com as publicações científicas, são, em muitos casos, insuficientes para garantir a reprodutibilidade e o reuso das contribuições científicas. Esta pesquisa objetiva definir uma estratégia computacional para o apoio ao reuso e a reprodutibilidade dos dados científicos, por meio da gestão da proveniência dos dados durante o seu ciclo de vida. A estratégia proposta nesta pesquisa é apoiada em dois componentes principais, um perfil de aplicação, que define um modelo padronizado para a descrição da proveniência dos dados, e uma arquitetura computacional para a gestão dos metadados de proveniência, que permite a descrição, armazenamento e compartilhamento destes metadados em ambientes distribuídos e heterogêneos. Foi desenvolvido um protótipo funcional para a realização de dois estudos de caso que consideraram a gestão dos metadados de proveniência de experimentos de modelagem de distribuição de espécies. Estes estudos de caso possibilitaram a validação da estratégia computacional proposta na pesquisa, demonstrando o seu potencial no apoio à gestão de dados científicos.

Palavras-chave: Proveniência dos Dados. Ciência dos Dados. Metadados. Pesquisas Reproduzíveis. Reuso de Dados

ABSTRACT

Modern science, supported by e-science, has faced challenges in dealing with the large volume and variety of data generated primarily by technological advances in the processes of collecting and processing scientific data. Therefore, there was also an increase in the complexity of the analysis and experimentation processes. These processes currently involve multiple data sources and numerous activities performed by geographically distributed research groups, which must be understood, reused and reproducible. However, initiatives by the scientific community with the goal of developing tools and sensitize researchers to share their data and source codes related to their findings, along with scientific publications, are often insufficient to ensure the reproducibility and reuse of scientific results. This research aims to define a computational strategy to support the reuse and reproducibility of scientific data through data provenance management during its entire life cycle. Two principal components support our strategy in this research, an application profile that defines a standardized model for the description of provenance metadata, and a computational architecture for the management of the provenance metadata that enables the description, storage and sharing of these metadata in distributed and heterogeneous environments. We developed a functional prototype for the accomplishment of two case studies that considered the management of provenance metadata during the experiments of species distribution modeling. These case studies enabled the validation of the computational strategy proposed in the research, demonstrating the potential of this strategy in supporting the management of scientific data.

Keywords: Data Provenance. Data Science. Metadata. Reproducible Research. Data Reuse.

LISTA DE FIGURAS

Figura 1 – Espectro da reprodutibilidade de pesquisas.....	21
Figura 2 - Dados, informação e conhecimento.....	33
Figura 3 – Representação da entropia de dados científicos.	38
Figura 4 - Esquema apresentando o Long Tail dos dados científicos.....	40
Figura 5 – Representação do antigo ciclo de vida dos dados.....	41
Figura 6 – Modelo de ciclo de vida dos dados científicos do DataONE.	42
Figura 7 – Modelo de ciclo de vida dos dados científicos descrito em Michener et al. (2011).....	43
Figura 8 – Modelo do DCAP segundo o Singapore Framework.....	51
Figura 9 – Camadas da Web Semântica.....	56
Figura 10 - Representação de uma afirmação em RDF, descrita por uma tripla.....	57
Figura 11 – Grafo RDF descrevendo um recurso.....	58
Figura 12 - Grafo RDF em Turtle, referente as triplas apresentadas na Figura 11. ...	59
Figura 13 – Demonstração do uso do SPARQL para a realização de uma consulta em um grafo RDF.....	60
Figura 14 – As três classes principais do PROV e suas relações.....	66
Figura 15 – Descrição em alto nível de um processo de monitoramento ambiental, por meio dos elementos fundamentais do modelo de dados do PROV.	67
Figura 16 – Esquema representando as informações necessárias para descrever um experimento científico.	74

Figura 17 – Modelo de domínio do Perfil de Aplicação para proveniência de dados de Biodiversidade.....	76
Figura 18 – Modelo de domínio do Perfil de Aplicação para proveniência de dados de Biodiversidade.....	77
Figura 19 - Pontos de Vista do RM-ODP.	82
Figura 20 – Representação dos componentes da arquitetura computacional e suas relações.....	91
Figura 21 – Exemplo de grafo RDF representado graficamente e no formato RDF/XML.	94
Figura 22 – Representação da distribuição dos componentes da arquitetura computacional proposta.	96
Figura 23 – Documento JSON-LD descrevendo a proveniência de um conjunto de dados.	97
Figura 24 – Documento de contexto, que descreve semanticamente os termos utilizados em um documento JSON-LD.	98
Figura 25 – Metadados após sua conversão pelo servidor de proveniência.....	99
Figura 26 – Representação dos componentes do protótipo funcional da arquitetura computacional de proveniência.	100
Figura 27 – Etapas que compõem o processo de modelagem de distribuição de espécies.	104
Figura 28 – Diagrama do modelo de dados de proveniência considerado no estudo de caso 1.....	107
Figura 29 – Etapas do estudo de caso 1.....	108
Figura 30 – Descrição dos metadados dos pesquisadores envolvidos no experimento.	109

Figura 31 – Descrição da atividade de modelagem do experimento.....	109
Figura 32 – Representação do grafo de proveniência da atividade de modelagem do experimento.....	110
Figura 33 – Função do pacote BioProv Client utilizada na requisição de dois relatórios sobre a proveniência do experimento.	111
Figura 34 – Relatório de proveniência do mapa de distribuição da Ziziphus Juazeiro gerado no navegador Web.....	111
Figura 35 - Relatório de proveniência do mapa de distribuição da Ziziphus Juazeiro gerado no formato PNG.	112
Figura 36 – À esquerda, a página Web no qual o mapa do modelo de distribuição foi publicado. À direita, o relatório de proveniência gerado pelo plug-in BioProv.....	113
Figura 37 – Metadados incluídos na página Web para identificar a existência de um recurso com proveniência disponível para consulta.....	113
Figura 38 – Etapas do processo do estudo de caso 2.	117
Figura 39 – Relatório de proveniência do experimento realizado no estudo de caso 2.	118
Figura 40 – Consulta SPARQL e sua requisição ao SPARQL endpoint do Servidor de Proveniência.....	119
Figura 41 – Consulta SPARQL da questão 1, do estudo de caso 2.....	119
Figura 42 – Consulta SPARQL da questão 2, do estudo de caso 2.....	120
Figura 43 – Consulta SPARQL da questão 3, do estudo de caso 2.....	121
Figura 44 – Consulta SPARQL da questão 4, do estudo de caso 2.....	122
Figura 45 – Consulta SPARQL da questão 5, do estudo de caso 2.....	124

LISTA DE TABELAS

Tabela 1 - Vocabulário de tipos do DCMI, contendo o nome do termo, sua URI associada e uma definição do termo.....	48
Tabela 2 - Lista de esquemas de metadados considerados em bases de dados de biodiversidade.	55
Tabela 3 – Mapeamento dos conceitos fundamentais do PROV.	65
Tabela 4 – Vocabulários considerados no PA proposto, com os respectivos namespaces e prefixos.....	79
Tabela 5 – Descrição da entidade Experimento Científico no DSP do perfil de aplicação BioProv.....	79
Tabela 6 – Descrição dos endpoints dos serviços de proveniência da arquitetura proposta.	92
Tabela 7 – Resultado da consulta SPARQL da questão 1, do estudo de caso 2....	120
Tabela 8 – Resultado da consulta SPARQL da questão 2, do estudo de caso 2....	121
Tabela 9 – Resultado da consulta SPARQL da questão 3, do estudo de caso 2....	122
Tabela 10 – Resultado da consulta SPARQL da questão 4, do estudo de caso 2..	123
Tabela 11 – Resultado da consulta SPARQL da questão 5, do estudo de caso 2..	124

LISTA DE ABREVIATURAS E SIGLAS

API	<i>Application Programming Interface</i>
ARK	<i>Archival Resource Key</i>
BI	<i>Biodiversity Informatics</i>
BIOPROV	<i>Biodiversity Provenance</i>
COL	<i>Catalogue of Life</i>
CoreCS	<i>Core Scientific Concepts</i>
CSDGM	<i>Content Standard for Digital Geospatial Metadata</i>
DataONE	<i>Data Observation Network for Earth</i>
DC	<i>Dublin Core</i>
DCMI	<i>Dublin Core Metadata Initiative</i>
DWC	<i>Darwin Core</i>
DOI	<i>Digital Object Identifier</i>
DSP	<i>Description Set Profile</i>
EML	<i>Ecological Metadata Language</i>
EXPO	<i>Ontology of Scientific Experiments</i>
FGDC	<i>Federal Geographic Documentation Committee</i>
FOAF	<i>Friend of a Friend Vocabulary</i>
GBIF	<i>Global Biodiversity Information Facility</i>
HTML	<i>HyperText Markup Language</i>
HTTP	<i>HyperText Transfer Protocol</i>
IB	<i>Informática para Biodiversidade</i>
ICMBIO	<i>Instituto Chico Mendes de Conservação da Biodiversidade</i>
ISO	<i>International Organization for Standardization</i>
ITIS	<i>Integrated Taxonomic Information System</i>
JBRJ	<i>Jardim Botânico do Rio de Janeiro</i>
JSON	<i>JavaScript Object Notation</i>
JSON-LD	<i>JavaScript Object Notation for Linked Data</i>
KNB	<i>Knowledge Network for Biocomplexity</i>
LSID	<i>Life Science Identifier</i>

MaxEnt	<i>Maximum Entropy</i>
MDE	<i>Modelagem de Distribuição de Espécies</i>
MMA	<i>Ministério do Meio Ambiente</i>
NISO	<i>National Information Standards Organization</i>
NSF	<i>National Science Foundation</i>
OPM	<i>Open Provenance Model</i>
OvO	<i>Open Provenance Ontology</i>
OWL	<i>Web Ontology Language</i>
PA	<i>Perfil de Aplicação</i>
PAV	<i>Provenance, Authoring and Versioning</i>
PNG	<i>Portable Network Graphics</i>
PROV	<i>Provenance Data Model</i>
RDF	<i>Resource Description Framework</i>
RF	<i>Requisito Funcional</i>
RFC	<i>Request for Comments</i>
RN	<i>Requisito Não Funcional</i>
REST	<i>Representational State Transfer</i>
RM-ODP	<i>Reference Model of Open Distributed Processing</i>
SIBBR	<i>Sistema de Informação sobre a Biodiversidade Brasileira</i>
SPARQL	<i>SPARQL Protocol and RDF Query Language</i>
SVG	<i>Scalable Vector Graphics</i>
TDWG	<i>Taxonomic Scientific Working Group</i>
Turtle	<i>Terse RDF Triple Language</i>
URI	<i>Uniform Resource Identifier</i>
URL	<i>Uniform Resource Locator</i>
URN	<i>Uniform Resource Name</i>
WfMS	<i>Workflow Management System</i>
W3C	<i>World Wide Web Consortium</i>
WWW	<i>World Wide Web</i>
XML	<i>eXtensible Markup Language</i>

SUMÁRIO

1	INTRODUÇÃO.....	19
1.1	MOTIVAÇÃO	22
1.2	CARACTERIZAÇÃO DO PROBLEMA	25
1.3	OBJETIVOS	25
1.4	MATERIAL E MÉTODO.....	26
1.5	ORGANIZAÇÃO DO TRABALHO.....	28
2	ASPECTOS CONCEITUAIS.....	30
2.1	INFORMÁTICA PARA BIODIVERSIDADE	30
2.2	GESTÃO DE DADOS	32
2.2.1	<i>Definição de Dados, Informação e Conhecimento</i>	<i>33</i>
2.2.2	<i>Definição de Gestão de Dados.....</i>	<i>34</i>
2.2.3	<i>Gestão de Dados Científicos.....</i>	<i>35</i>
2.2.4	<i>Ciclo de Vida dos Dados</i>	<i>40</i>
2.3	METADADOS	44
2.3.1	<i>Vocabulários.....</i>	<i>46</i>
2.3.2	<i>Vocabulário Controlado.....</i>	<i>47</i>
2.3.3	<i>Perfil de Aplicação.....</i>	<i>49</i>
2.3.4	<i>Sintaxe dos Metadados.....</i>	<i>51</i>
2.3.5	<i>Vocabulários na área de Biodiversidade</i>	<i>52</i>

2.3.6	<i>Web Semântica</i>	55
2.4	PROVENIÊNCIA DOS DADOS.....	61
2.4.1	<i>Conceitos Gerais</i>	61
2.4.2	<i>Modelos de proveniência dos dados</i>	62
2.4.3	<i>Soluções computacionais para a Proveniência de Dados</i>	68
3	PERFIL DE APLICAÇÃO PARA PROVENIÊNCIA DE DADOS DE BIODIVERSIDADE	72
3.1	PROCESSO DE DESENVOLVIMENTO DO PERFIL DE APLICAÇÃO PARA PROVENIÊNCIA DE DADOS DE BIODIVERSIDADE.....	73
3.1.1	<i>Especificação dos Requisitos Funcionais</i>	73
3.1.2	<i>Desenvolvimento do Modelo de Domínio</i>	76
3.1.3	<i>Desenvolvimento do Description Set Profile (DSP)</i>	78
4	ARQUITETURA COMPUTACIONAL PARA GESTÃO DE METADADOS DE PROVENIÊNCIA.....	81
4.1	REQUISITOS PARA A PROVENIÊNCIA DOS DADOS.....	82
4.1.1	<i>Requisitos de Conteúdo</i>	83
4.1.2	<i>Requisitos de Gerenciamento</i>	85
4.1.3	<i>Requisitos de Uso</i>	87
4.2	PONTOS DE VISTA DO RM-ODP	89
4.2.1	<i>RM-ODP: Ponto de Vista Empresarial</i>	89
4.2.2	<i>RM-ODP: Ponto de Vista de Informação</i>	90
4.2.3	<i>RM-ODP: Ponto de Vista Computacional</i>	90
4.2.4	<i>RM-ODP: Ponto de Vista de Engenharia</i>	95

4.2.5 RM-ODP: Ponto de Vista de Tecnologia	99
5 VALIDAÇÃO DA PROPOSTA	103
5.1 PROCESSO DE MODELAGEM DE DISTRIBUIÇÃO DE ESPÉCIES	103
5.2 ESTUDO DE CASO 1	106
5.2.1 Contextualização	106
5.2.2 Descrição do experimento	107
5.2.3 Discussões	113
5.3 ESTUDO DE CASO 2	115
5.3.1 Contextualização	115
5.3.2 Descrição do Experimento	116
5.3.3 Discussões	124
5.4 CONSIDERAÇÕES FINAIS	125
6 CONCLUSÃO	125
6.1 CONTRIBUIÇÕES DA TESE	130
6.2 TRABALHOS FUTUROS	131
7 REFERÊNCIAS	133
APÊNDICE A – DESCRIPTION SET PROFILE (DSP) DO PERFIL DE APLICAÇÃO PROPOSTO E VOCABULÁRIOS CONTROLADOS	147
APÊNDICE B – REPOSITÓRIOS DE CÓDIGOS-FONTE DOS COMPONENTES DO PROTÓTIPO FUNCIONAL	153
APÊNDICE C – SCRIPT DO EXPERIMENTO REALIZADO NO ESTUDO DE CASO 1	154
APÊNDICE D – RELATÓRIO DE PROVENIÊNCIA DO ESTUDO DE CASO 1 (FORMATO PNG)	161

APÊNDICE E – TRECHO DO RELATÓRIO DE PROVENIÊNCIA DO ESTUDO DE CASO 1 (FORMATO RDF/TURTLE).	162
APÊNDICE F – SCRIPT DO EXPERIMENTO REALIZADO NO ESTUDO DE CASO 2.	163
APÊNDICE G – RELATÓRIO DE PROVENIÊNCIA DO ESTUDO DE CASO 2 (FORMATO PNG).	176
APÊNDICE H – TRECHO DO RELATÓRIO DE PROVENIÊNCIA DO ESTUDO DE CASO 2 (FORMATO RDF/TURTLE).	177

1 INTRODUÇÃO

A ciência está em constante evolução. As descobertas científicas geradas em um único laboratório, por um ou poucos cientistas, em um único departamento, instituto ou universidade estão tornando-se cada vez mais raras. As principais questões científicas necessitam da aplicação do conhecimento, das habilidades e dos recursos de múltiplos pesquisadores e instituições.

A ciência também apoia-se cada vez mais nos dados. Novas tecnologias aumentaram a eficiência na coleta de dados e, conseqüentemente, a quantidade dos dados gerados. As tecnologias também permitiram que os dados pudessem ser processados, analisados e armazenados em infraestruturas computacionais. Este cenário gerou uma mudança nos desafios científicos. O antigo problema da escassez dos dados, foi substituído pela dificuldade em gerenciar o seu excesso, sua variedade e distribuição (HEY; TANSLEY; TOLLE, 2009).

Uma nova área científica nasceu dentro desta realidade. Apesar dos diferentes nomes usados, entre eles *e-science*, *data-driven science* ou *data-intensive science*, nesta nova ciência, os cientistas e engenheiros da computação uniram-se a especialistas de diferentes domínios para o desenvolvimento de teorias e técnicas, a partir da massiva quantidade de dados disponibilizados pelas áreas do conhecimento (HEY; TANSLEY; TOLLE, 2009). A área de *e-science* contempla tanto as questões de *hardware* (sensores, sistemas computacionais de armazenamento, processamento e comunicação, dispositivos de visualização, etc.) como de *software*, para apoiar as diversas etapas envolvidas no processo científico, como a aquisição, processamento, armazenamento, análise e visualização dos dados, considerando assim, todo o ciclo de vida dos dados (MICHENER; JONES, 2012).

Os desafios que influenciaram o surgimento da *e-science*, também geraram mudanças nas disciplinas tradicionais, que criaram ramos computacionais para lidar com a gestão de dados dentro de cada área científica. Na biodiversidade, por exemplo, surgiu a Informática para Biodiversidade (IB), na área da biologia molecular, mais especificamente do genoma, surgiu a Bioinformática, na área da geografia surgiu

a geoinformática, dentre outras. De maneira geral, esse tema vem se consolidando numa área denominada ciência dos dados.

A ciência dos dados representa o campo de estudo interdisciplinar que utiliza técnicas de diversas áreas científicas para definir processos e sistemas para a extração de informação e conhecimento a partir dos dados (DHAR, 2013; PROVOST; FAWCETT, 2013).

Juntamente com os desafios de lidar com grande volume e variedade de dados, as iniciativas para a gestão dos dados científicos também foram intensificadas devido ao aumento na complexidade dos processos de análise e experimentação. Estes processos consideram múltiplas fontes de dados e atividades, além de serem realizadas por grupos de pesquisadores geograficamente distribuídos. É essencial que estes dados e processos possam ser compreendidos, reutilizados e reproduzíveis (DEELMAN et al., 2010).

Conforme descrito em Stodden et al. (2014), a ciência avança quando as descobertas são replicadas e reproduzidas. A replicação, pode ser descrita como a prática de implementar, de forma independente, experimentos científicos para a avaliação das suas descobertas, e é a base da validação e da confiabilidade das descobertas científicas. Relacionado a replicação, existe a reprodução, que pode ser definida como uma variação da replicação, na qual não é realizada uma nova coleta de dados e são considerados os mesmos métodos e dados da pesquisa original (PENG, 2009; PENG; DOMINICI; ZEGGER, 2006). Na literatura, geralmente a replicabilidade e a reprodutibilidade são considerados sinônimos ou possuem pequenas variações nas suas definições. Neste trabalho, iremos considerar a reprodutibilidade como a característica de uma pesquisa ser compreendida, validada e refeita, por pesquisadores independentes, a partir das informações disponibilizadas sobre o processo e os dados originais.

O reuso dos dados publicados para a geração de novos conhecimentos também é essencial para a evolução da ciência (FREIRE et al., 2014). Além de possibilitar a reanálise de evidências, a minimização da duplicação de esforços e a aceleração da inovação. No entanto, assim como na reprodutibilidade dos experimentos, o reuso dos

resultados depende do entendimento e da avaliação do processo que gerou esses dados, garantindo assim, a sua confiabilidade e sua adequabilidade aos novos processos que irão consumir estes dados.

Dentro deste contexto, o conhecimento compartilhado por meio da prática padrão das publicações de artigos científicos geralmente não é suficiente para a reprodutibilidade dos experimentos e o reuso dos seus produtos (MESIROV, 2010; PENG, 2011). Os pesquisadores precisam confiar em tabelas, gráficos e legendas de figuras, incluídas nas publicações, por não possuírem informações suficientes para a sua reprodução. Consequentemente, torna-se um grande desafio a verificação e a reprodução dos resultados publicados, tanto para os revisores da pesquisa quanto para os demais pesquisadores (LEVEQUE, 2009). Esta situação ocasionou a chamada crise de credibilidade nas ciências (BAKER, 2016; DONOHO et al., 2009).

A pesquisa reproduzível passou a ser um tema de pesquisa e um objetivo a ser alcançado pelas instituições e pesquisadores (PENG, 2009, 2011; VANDEWALLE; KOVACEVIC; VETTERLI, 2009). Estas iniciativas buscam o aumento no grau de reprodutibilidade das descobertas científicas. Em Peng (2011) é apresentado o espectro da reprodutibilidade (Figura 1), que mostra que uma pesquisa pode ter níveis de reprodutibilidade diferenciados, com base no nível de compartilhamento dos dados e códigos utilizados na pesquisa.

Figura 1 – Espectro da reprodutibilidade de pesquisas.

Fonte: Adaptado de (PENG, 2011).

No entanto, existem pesquisas na qual a disponibilização dos dados e dos códigos ainda são insuficientes para garantir a reprodutibilidade e o reuso das descobertas científicas. É necessário o compartilhamento de informações adicionais sobre a pesquisa, incluindo, equipamentos e ferramentas utilizadas, dados brutos, processos e as decisões consideradas na geração deste conhecimento. Assim, este trabalho propõe uma estratégia computacional, baseada na proveniência de dados, para a gestão das informações sobre os processos e dados gerados durante um experimento científico, buscando auxiliar na disponibilização das informações necessárias para tornar as descobertas científicas reproduzíveis e reutilizáveis por outros pesquisadores.

1.1 Motivação

No contexto da biodiversidade, o surgimento da IB, do inglês *Biodiversity Informatics*, foi motivado pela necessidade de se gerenciar o crescente volume de dados gerados pelas pesquisas, garantindo que estes dados pudessem ser descobertos, integrados e reutilizados por outros pesquisadores (PAGE, 2008; SOBERÓN; PETERSON, 2004).

Nos últimos anos, a comunidade internacional de IB intensificou os esforços para aprimorar os processos de gerenciamento dos dados científicos. Vários avanços foram alcançados, principalmente na padronização, na preservação de dados a longo prazo e no compartilhamento de dados de biodiversidade para o seu uso em questões críticas, como a conservação e uso sustentável do meio ambiente, planos de ação para mudanças climáticas, e serviços ecossistêmicos (BACH et al., 2012; CANDELA et al., 2014; HEIDORN, 2011; JOHNSON, 2007; WALLIS; ROLANDO; BORGMAN, 2013).

A aplicação de padrões de metadados e protocolos de comunicação foram as bases que permitiram estes avanços. Dentre estes padrões de metadados, destacaram-se o Darwin Core (WIECZOREK et al., 2012), usado para descrever a ocorrência de vida no planeta e suas associações com o meio ambiente; o *Ecological*

Metadata Language (EML) (FEGRAUS et al., 2005), usado para descrever dados ecológicos; e o Dublin Core (WEIBEL et al., 1998), usado para descrever recursos digitais. Os padrões de metadados foram utilizados principalmente com o objetivo de descrever e estruturar, de forma padronizada, os dados gerados pelas diversas iniciativas, facilitando a sua descoberta e sua interoperabilidade entre diferentes instituições, sistemas e bases de dados.

No entanto, estes padrões de metadados são formados, na sua maioria, de elementos descritivos de um recurso de dados, utilizados para facilitar a sua descoberta e identificação (NISO, 2004). Os elementos dos metadados que tratam sobre a origem dos dados, seu processo de geração e questões administrativas, como políticas de uso, são raramente descritos.

Por isso, para o reuso dos dados disponíveis em novas pesquisas e em processos de tomada de decisão, pesquisadores e gestores precisam de informações adicionais relacionadas a origem destes dados, quais os métodos utilizados na sua geração, quem são os responsáveis pela pesquisa, além de informações técnicas que possam garantir a confiabilidade e a validade destes dados (ENKE et al., 2012; NISO, 2004).

Para as descobertas científicas, geradas nos projetos e experimentos, o problema intensifica-se, pois, muitos dos resultados, são descritos apenas por meio de artigos científicos, que geralmente disponibilizam informações insuficientes para garantir o reuso dos dados e a reprodutibilidade destes experimentos (CASSEY; BLACKBURN, 2006; COSTELLO et al., 2013; COSTELLO; WIECZOREK, 2014).

A busca de uma estratégia que possa descrever e armazenar as informações vitais sobre os processos geradores dos dados, e sua posterior disponibilização de forma aberta na Web, para apoiar o reuso destes dados, é um fator motivador da tese.

As informações que descrevem os processos envolvidos na geração e evolução dos dados científicos também são chamadas de proveniência dos dados. A proveniência dos dados, dentro do contexto da gestão de dados científicos, pode ser usada para a avaliação da qualidade e confiabilidade dos dados, na auditoria de

processos, na reprodutibilidade de experimentos, na descoberta de novos dados, entre outros usos (SIMMHAN; PLALE; GANNON, 2005). No entanto, para a obtenção dos benefícios da proveniência dos dados, os cientistas precisam gerenciar e manter estes metadados durante todas as etapas que compõem o ciclo de vida dos dados.

Grande parte das soluções para proveniência de dados científicos são baseadas nos sistemas gerenciadores de *workflow*, do inglês *Workflow Management System* (WfMS) (DAVIDSON; FREIRE, 2008). No entanto, a área de biodiversidade possui uma vasta gama de aplicações e ferramentas que são utilizadas durante todo o ciclo de vida dos dados, o que dificulta a implantação de soluções que exijam a utilização de ferramentas específicas. Além disso, questões como aprendizado, confiança em ferramentas conhecidas e outros fatores não tecnológicos também são barreiras que dificultam a inserção de novas ferramentas computacionais.

A busca por uma solução para a gestão da proveniência dos dados que seja independente de ferramentas específicas é um motivador da tese.

Por fim, a participação em projetos de pesquisa, como o Herbário Virtual REFLORE (<http://reflora.jbrj.gov.br>) (FORZZA et al., 2016), o Portal da Biodiversidade do Ministério do Meio Ambiente (MMA) (<http://portaldabiodiversidade.icmbio.gov.br>) e o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBR) (GADELHA et al., 2014), possibilitaram o entendimento dos reais problemas de gestão de dados existentes no cenário nacional e internacional da área de IB.

Adicionalmente, com a organização e participação de eventos científicos, como o *8th International Conference on Ecological Informatics* (ISEI) e o *Workshop em Ciência dos Dados* (WDS), a participação no Grupo de Trabalho de Integração de Dados do MMA (CORRÊA, 2013), além da escrita de um livro sobre diretrizes técnicas para a integração de dados de biodiversidade (SILVA et al., 2015), foi possível a realização de discussões com pesquisadores da área de IB e o entendimento sobre a riqueza das análises e dados existentes no área. Estes processos e dados necessitam de uma documentação detalhada para serem compreendidos e reutilizados.

Estas experiências, vivenciadas durante a pesquisa, também foram motivadores para esta tese.

1.2 Caracterização do Problema

A publicação de pesquisas reproduzíveis é um dos principais objetivos da ciência. No entanto, com o surgimento da *e-science*, caracterizada pela quantidade, variedade e complexidade dos dados, ferramentas e processos, é necessário a disponibilização de informações adicionais, juntamente com a publicação das descobertas científicas, que garantam que outros pesquisadores possam avaliar, reutilizar e reproduzir estes resultados. Nesse sentido, *como definir uma estratégia computacional, alinhada ao ciclo de vida dos dados científicos, que possibilite: (i) o registro dos dados necessários para garantir a reprodutibilidade e a reuso das descobertas científicas e (ii) a disponibilização destes dados, de forma aberta na Web, juntamente com a publicação dos resultados da pesquisa?* Esse questionamento deu origem a tese de que a gestão da proveniência dos dados científicos e sua disponibilização na Web possa auxiliar na geração de pesquisas reproduzíveis e reutilizáveis.

1.3 Objetivos

O objetivo principal da tese é definir uma estratégia computacional, baseada na gestão e compartilhamento dos metadados de proveniência, capturados durante todo o ciclo de vida dos dados, para apoiar a reprodutibilidade de experimentos na área de biodiversidade e a reutilização dos dados científicos gerados.

Os seguintes objetivos específicos foram estabelecidos para a pesquisa:

- Definir um perfil de aplicação, baseado em ontologias e vocabulários existentes, para a estruturação, padronização e armazenamento dos metadados de proveniência;

- Definir uma arquitetura computacional para a captura, gestão e compartilhamento na Web dos metadados de proveniência de recursos de dados e experimentos científicos, garantindo assim, que estes dados possam ser avaliados e reutilizados pelos pesquisadores;
- Validar o perfil de aplicação e a arquitetura computacional para gestão de metadados de proveniência em estudos de caso com cenários na área de biodiversidade.

1.4 Material e Método

Esta pesquisa considerou uma abordagem exploratória e de natureza qualitativa (WAZLAWICK, 2009).

A pesquisa foi iniciada com uma revisão bibliográfica exploratória da área de IB (HARDISTY; ROBERTS, 2013; JOHNSON, 2007), com foco na aplicação de protocolos e padrões de metadados, como o *TDWG Access Protocol for Information Retrieval* (TAPIR) (BACH et al., 2012), o Darwin Core (BAKER; RYCROFT; SMITH, 2014; WIECZOREK et al., 2012), o EML (HARRINGTON, 2011) e o Dublin Core (WEIBEL et al., 1998), para apoiar os processos de padronização e integração de sistemas e bases de dados de biodiversidade. Com o entendimento destes desafios, novos temas foram abordados na pesquisa, como a gestão de dados científicos e o ciclo de vida dos dados. Entre os desafios para a gestão dos dados da área de IB, a reutilização dos dados científicos foi escolhida para a continuidade da pesquisa.

Uma nova revisão bibliográfica exploratória foi realizada sobre gestão de dados científicos (COSTELLO; WIECZOREK, 2014; DAMA INTERNATIONAL, 2009; RÜEGG et al., 2014), pesquisas reproduzíveis (PENG, 2011) e a proveniência de dados (FREIRE et al., 2008), com foco nos modelos de proveniência (BELHAJJAME et al., 2013) e nas arquiteturas computacionais existentes (SIMMHAN; PLALE; GANNON, 2005). A partir deste conhecimento, foi possível avaliar a estratégia mais adequada para a pesquisa.

Após definido o escopo e os objetivos da pesquisa, o procedimento metodológico considerou 5 etapas:

- **Estudo do domínio do problema:** Revisão bibliográfica sobre metadados (HASLHOFER; KLAS, 2010), ontologias (PAGE, 2006) e perfis de aplicação (MALTA; BAPTISTA, 2014), com o objetivo de definir um modelo de dados padronizado para proveniência de dados científicos. Também foi realizada uma revisão bibliográfica sobre a arquitetura computacional da Web Semântica (ANTONIOU; ANTONIOU, 2012), com foco no *Resource Description Framework* (RDF) (CYGANIAK; WOOD; LANTHALER, 2014) e no SPARQL (THE W3C SPARQL WORKING GROUP, 2013), com o objetivo destas tecnologias serem consideradas na arquitetura computacional para a gestão dos metadados de proveniência.
- **Especificação da Arquitetura Computacional:** Especificação da arquitetura computacional para a gestão dos metadados de proveniência em ambientes distribuídos e heterogêneos. A arquitetura é baseada em serviços Web que consideram o estilo arquitetural REST (FIELDING, 2000) e a serialização dos dados em JSON-LD (SPORNY et al., 2014), os metadados são organizados e armazenados em RDF e a recuperação dos dados utiliza o SPARQL. As decisões arquiteturais foram baseadas nos requisitos funcionais e não funcionais descritos no Capítulo 4. A descrição da arquitetura computacional foi baseada no Modelo de Referência para Processamento Distribuído Aberto, do inglês *Reference Model of Open Distributed Processing* (RM-ODP) (ISO/IEC 10746-3:2009, 2009).
- **Definição do Modelo de Dados:** Definição de perfil de aplicação baseado no W3C PROV (BELHAJJAME et al., 2013) e em dicionários existentes, como o Dublin Core (WEIBEL et al., 1998), FOAF (BRICKLEY; MILLER, 2014) e Darwin Core (WIECZOREK et al., 2012), para possibilitar a padronização na descrição dos metadados de proveniência;

- **Construção de um Protótipo Funcional:** Implementação do protótipo funcional, baseado na arquitetura computacional especificada, utilizando tecnologias como o Python, Javascript e R, além de bancos de dados PostgreSQL e AllegroGraph.
- **Execução de Estudos de Caso:** Execução de dois estudos de caso para a validação da estratégia computacional proposta. Estes estudos de caso consideraram a execução de experimentos de modelagem de distribuição de espécies (MDE) (CORRÊA et al., 2011). O primeiro experimento registrou a proveniência do modelo de distribuição da espécie *Ziziphus Juazeiro* na região semiárida brasileira. O segundo, registrou a proveniência do modelo de distribuição da espécie *Urocitellus beldingi* na região da Califórnia, nos Estados Unidos. Nestes estudos de caso o perfil de aplicação e a arquitetura computacional foram validados. Enquanto o primeiro estudo de caso foi direcionado para a validação dos componentes da arquitetura computacional, o segundo validou o perfil de aplicação, além da análise e extração dos metadados de proveniência por meio da submissão de consultas SPARQL.

1.5 Organização do Trabalho

Esta tese possui, além da Introdução, 6 capítulos, os quais são organizados conforme descritos a seguir.

O Capítulo 2 discute os aspectos conceituais relacionados ao tema da pesquisa. Inicia com uma introdução sobre a área de IB. Em seguida descreve os conceitos de gestão de dados e ciclo de vida dos dados. Também são apresentados os conceitos de metadados, seus padrões e sua evolução com o surgimento da Web Semântica. Por fim, são apresentados os conceitos de proveniência de dados científicos, os modelos de dados existentes e o estado da arte da proveniência na área científica.

O Capítulo 3 descreve a definição de um perfil de aplicação para a proveniência de dados científicos. A partir da utilização do *Singapore Framework*, é definido um perfil de aplicação para a descrição da proveniência dos dados em experimentos científicos.

O Capítulo 4 descreve uma arquitetura computacional para a gestão de metadados de proveniência durante o ciclo de vida dos dados. Em seguida apresenta o protótipo funcional desenvolvido para a validação da arquitetura proposta. A documentação desta arquitetura é baseada no padrão RM-ODP.

O Capítulo 5 apresenta dois estudos de caso, no qual o protótipo desenvolvido foi utilizado para a gestão da proveniência dos dados gerados em experimentos de modelagem de distribuição de espécies. Os resultados dos estudos de caso são apresentados e discutidos.

O Capítulo 6 apresenta as conclusões do trabalho, as principais contribuições e as dificuldades encontradas durante a pesquisa. Também são descritos os trabalhos em andamento e os trabalhos futuros relacionados ao tema.

Concluindo, são apresentadas as referências utilizadas na elaboração desta pesquisa e os apêndices que fazem parte deste trabalho.

2 ASPECTOS CONCEITUAIS

2.1 Informática para Biodiversidade

A biodiversidade pode ser definida como a variabilidade entre os organismos vivos de todas as origens, tais como os ecossistemas terrestre, marinho, outros ecossistemas aquáticos e os complexos ecológicos de que dependem, incluindo a diversidade dentro da espécie, entre espécies e entre os ecossistemas (ALCAMO et al., 2003).

As informações da biodiversidade são utilizadas em uma ampla variedade de atividades nas áreas científicas, educacionais e governamentais, e são críticas em processos de tomada de decisão em diversas domínios (CANHOS et al., 2004). Por isso, existe uma demanda crescente na compreensão e solução dos problemas complexos que envolvem o meio ambiente (CANHOS; SARAIVA, 2012).

Para buscar soluções para estes problemas, é essencial a realização de análises abrangentes, a partir do uso de informações integradas sobre a biodiversidade, que contemplem um amplo escopo espacial e temporal e que sejam provenientes de diferentes fontes de dados. Estas informações também precisam estar associados a dados adicionais sobre ecossistemas, mudanças climáticas, dados abióticos, entre outros (CANHOS et al., 2004).

Neste contexto, a IB foi criada. A IB é uma recente área de pesquisa, com a sua primeira citação em 1998 (SCHALK, 1998), e tem como objetivo a aplicação das técnicas de computação na área da biodiversidade, para possibilitar melhorias na descoberta, gerenciamento, exploração, apresentação e análise destes dados (SOBERÓN; PETERSON, 2004).

As informações sobre biodiversidade englobam uma variedade de tipos de dados, tais como, dados morfológicos, dados genômicos, dados geoespaciais, dados taxonômicos, dados nomenclaturais, dados ambientais, entre outros. Desta quantidade de tipos de dados, os dados primários de biodiversidade são os mais populares na comunidade de IB, devido a sua aplicação em uma ampla variedade de estudos (CHAPMAN, 2005a).

Os dados primários descrevem a ocorrência de um táxon específico na natureza, considerando as dimensões espacial (o local que ocorreu) e temporal (quando ocorreu) (CHAPMAN, 2005b). A partir das análises, consolidações, interpretações e suposições realizadas sobre os dados primários e demais dados associados, são gerados seus resultados, os chamados dados secundários.

Segundo Peterson et al. (2010), o estágio atual da área de IB pode ser dividido em três categorias gerais:

- Extração e captura dos dados;
- Integração e disponibilização dos dados;
- Visualização e análise dos dados.

As iniciativas de extração e captura dos dados incluem a digitalização de amostras de espécimes e a triagem, por meio da digitalização da literatura, dos nomes científicos e localizações geográficas registradas (GURALNICK; HILL, 2009). A geração de dados por meios dos projetos de cidadão-cientista também se tornaram uma realidade (SULLIVAN et al., 2009), (DICKINSON; ZUCKERBERG; BONTER, 2010). As iniciativas de interoperabilidade concentram-se na organização dos dados e metadados gerados pelas instituições e pesquisadores, para consolidá-los e publicá-los em redes de informação de biodiversidade (ROBERTSON et al., 2014). Em paralelo, grandes esforços são investidos na compilação de dados taxonômicos extraídos das literaturas científicas (KONING; SARKAR; MORITZ, 2005). Por fim, a visualização eficiente destas informações incluem a apresentação dos dados de biodiversidade compiladas em visões analíticas e mapas ou em *mashups* com dados relacionados de diversas fontes (JAROENSUTASINEE; PHEERA; JAROENSUTASINEE, 2014), (BUTLER, 2006).

O resultado dos esforços da comunidade de IB nos últimos anos foi o surgimento e a consolidação de diversas iniciativas voltadas para a gestão e compartilhamento de dados de biodiversidade. Entre estas iniciativas podemos citar o *Global Biodiversity Information Facility* (GBIF) (BOAKES et al., 2010; YESSON et al., 2007), que atualmente indexa mais de 600 milhões de registros no formato Darwin Core, proveniente de mais de 800 organizações de 54 países. O DataONE (ALLARD, 2012;

MICHENER et al., 2012), programa patrocinado pela *National Science Foundation* (NSF) dentro do programa *DataNet*, com o objetivo principal de armazenar e preservar dados ecológicos e ambientais produzidos por cientistas do mundo inteiro. E o *Catalogue of Life* (CoL) (JONES et al., 2010; JONES; WHITE; ORME, 2011), criado a partir de uma parceria entre o programa *Species 2000* e o *Integrated Taxonomic Information System* (ITIS), com o objetivo de gerar um catálogo abrangendo de todas as espécies do planeta. Atualmente, este catálogo contempla uma compilação de 158 bases de dados taxonômicas.

No Brasil, podemos destacar o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr) (GADELHA et al., 2014), que indexa mais de 10 milhões de registros provenientes de instituições nacionais, o Portal da Biodiversidade do MMA, que indexa dois milhões de registros provenientes dos centros de pesquisa e conservação do Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) e do Jardim Botânico do Rio de Janeiro (JBRJ) . E o SpeciesLink, que disponibiliza mais de 8 milhões de registros provenientes de coleções biológicas nacionais e internacionais.

Estes últimos anos mostraram que a comunidade de IB alcançou vários avanços que geraram benefícios para pesquisadores, instituições, financiadores e para a sociedade em geral. Em paralelo aos desafios da conscientização dos pesquisadores na importância de gerenciar e compartilhar, de forma aberta, os dados científicos, estes avanços também estão possibilitando que novas problemáticas sejam analisados e, conseqüentemente, que a área continue evoluindo em busca de melhores processos, técnicas e ferramentas para apoiar o processo científico.

2.2 Gestão de Dados

Primeiramente, é importante definir dados, informação e conhecimento, que são termos relacionados, mas que possuem diferentes definições na literatura.

2.2.1 Definição de Dados, Informação e Conhecimento

Os dados são as representações de fatos na forma de textos, números, gráficos, imagens, sons ou vídeos. Fatos são capturados, armazenados e expressados como dados (DAMA INTERNATIONAL, 2009).

A informação são os dados em um contexto. Sem um contexto, os dados não possuem sentido. Uma informação significativa é criada a partir da interpretação do contexto relacionado aos dados. Este contexto deve incluir:

- O significado dos elementos e dos termos relacionados;
- O formato no qual os dados estão representados;
- O período representado pelos dados;
- Os objetivos buscados durante a geração destes dados;
- A relevância dos dados para um determinado uso.

Os dados são a matéria-prima que os consumidores de dados interpretam para continuamente gerar a informação, conforme mostrado na Figura 2. A informação resultante deste processo direciona nossas decisões.

Figura 2 - Dados, informação e conhecimento.

Fonte: Adaptado de (DAMA INTERNATIONAL, 2009).

As informações contribuem para o conhecimento. O conhecimento é a compreensão, consciência e o reconhecimento de uma situação. O conhecimento é a informação em uma perspectiva, integrado a um ponto de vista com base no reconhecimento e interpretação de padrões, tais como tendências, formadas a partir de outras informações e experiências. Podendo ser incluídas também hipóteses e teorias. O conhecimento é obtido quando o significado das informações é compreendido (DAMA INTERNATIONAL, 2009).

Os dados são a base da informação e do conhecimento, mas dados imprecisos, incompletos, desatualizados e incompreensíveis podem gerar falsas afirmações, que irão apoiar decisões incorretas. Por isso, a grande importância do reconhecimento da gestão de dados como parte fundamental dos processos realizadas nas organizações.

2.2.2 Definição de Gestão de Dados

A gestão de Dados, do inglês *Data Management*, é a disciplina responsável por definir, planejar, implantar e executar estratégias, procedimentos e práticas necessárias para a gestão efetiva dos recursos de dados e informações das organizações, incluindo planos para sua definição, padronização, organização, proteção e utilização (DAMA INTERNATIONAL, 2009). As atividades da gestão de dados são consideradas nas diversas áreas de uma instituição, estendendo-se aos parceiros, fornecedores e consumidores. No nível institucional, envolve desde os gestores, que utilizam dados para a tomada de decisões, até profissionais de nível operacional, que são responsáveis pela coleta, produção e análise dos dados.

Entre os principais objetivos da gestão de dados destacam-se (DAMA INTERNATIONAL, 2009; EYNDEN; UK DATA ARCHIVE, 2011; STRASSER et al., 2012):

- Garantia da qualidade dos dados;
- Correta utilização e reutilização dos dados;
- Decisões ágeis e corretas baseadas nos dados;
- Confiabilidade e proveniência dos dados;

- Segurança e gestão de riscos;
- Manutenção dos dados a longo prazo.

2.2.3 Gestão de Dados Científicos

Na área científica, a gestão de dados é a compilação de diversas boas práticas aplicadas ao processo de pesquisa que tornam os dados fáceis de serem descobertos, compreendidos, reutilizáveis e menos passíveis de perdas ou esquecimento (BRINEY, 2015).

Diversas são as motivações para a aplicação da gestão de dados em pesquisas científicas. Entre elas destacam-se (BORGMAN, 2012; DA SILVA; CORREA; STANZANI, 2015):

- **Aumento na visibilidade da pesquisa:** Tornando o dado visível para outros pesquisadores por meio de repositórios reconhecidos, é possível demonstrar a relevância da pesquisa, garantindo também o uso contínuo destas informações para outras finalidades. Além disso, com as técnicas de identificação de conjuntos de dados por identificadores globais, como o *Digital Object Identifier* (DOI) (SIMONS, 2012), sua utilização poderá ser rastreada e a sua relevância mensurada;
- **Economia de tempo e recursos:** A médio e longo prazo, o planejamento da gestão dos dados possibilita uma economia de tempo e de recursos. Por exemplo, novos projetos de pesquisa podem ser direcionados com base nos dados históricos conhecidos, evitando que esforços duplicados sejam realizados em áreas que possuem dados suficientes, passando a ser direcionados para áreas com lacunas de conhecimento ou pouco exploradas;
- **Facilidade na replicação e divulgação:** A disponibilização dos dados, com todos seus metadados, em um repositório acessível, possibilita que outros pesquisadores possam localizar, compreender e reutilizar os dados de forma mais eficiente, sem a necessidade de suporte do pesquisador que gerou os dados.

- **Preservação dos dados:** Ao depositar os dados em um repositório confiável, aumentam as chances destas informações permanecerem seguras e disponíveis no longo prazo. Com isso, o investimento de tempo e recursos é protegido, e a contribuição do pesquisador preservada.
- **Aumento na eficiência das pesquisas:** A gestão eficiente dos dados gerados em uma pesquisa possibilita que novas pesquisas possam ser realizadas a partir de dados existentes, sem a necessidade de novos esforços para a coleta e processamento de dados. A reutilização de dados existentes possibilita a economia de recursos e uma maior eficiência no processo científico.
- **Documentação dos dados:** Gerenciar e documentar os dados durante seu ciclo de vida garante que a integridade e a descrição adequada dos dados sejam mantidas. A tentativa de documentação de dados históricos mostrou-se ineficiente e custosa.
- **Atendimento aos requisitos das financiadoras:** É cada vez mais comum que as agências de financiamento exijam o depósito de todos os dados gerados na pesquisa como um produto do projeto financiado. Estes dados devem seguir normas, padrões e protocolos que garantem a preservação e a reutilização destes dados.
- **Facilidade em novas descobertas:** Possibilitar que outros pesquisadores possam avançar em partes inacabadas de uma pesquisa ou realizar descobertas inesperadas, inclusive em áreas diferentes da pesquisa que gerou os dados.
- **Apoio ao acesso aberto:** Os pesquisadores estão tornando-se cada vez mais conscientes da necessidade de gerenciar de forma eficiente os dados gerados e considerar as iniciativas de compartilhamento do conhecimento da comunidade científica. O movimento “*Open Data*” defende que pesquisadores precisam compartilhar seus dados, ampliando o desenvolvimento do conhecimento mundial.

No entanto, desafios atuais ainda dificultam a aplicação das boas práticas de gestão de dados nas instituições e nos grupos de pesquisa na área de biodiversidade:

- **Armazenamento em Silos de Dados:** Uma grande quantidade de dados é armazenada em depósitos isolados de dados, criados pelas organizações, seus setores e até pelos próprios pesquisadores e colaboradores. Apesar dos maiores problemas serem causados pelos silos de dados isolados, que não permitem o acesso aos dados, um outro desafio é gerado pelas próprias iniciativas da comunidade de IB, que geram grandes bases de dados de tipos de dados específicos. Estas bases de dados possuem baixa ou nenhuma integração com bases de dados voltadas a outros tipos de dados (PETERSON; SOBERÓN; KRISHTALKA, 2015). Com isso, o pesquisador ainda possui um grande esforço no processo de obtenção dos dados necessários para a sua pesquisa.
- **Dados órfãos:** Informações são perdidas em uma instituição em situações como a aposentadoria de um pesquisador, o roubo ou inutilização de um notebook que não teve seus dados copiados em um servidor seguro, ou no simples fato de uma informação importante ter sido esquecida em um meio de armazenamento inacessível, como computador, mídia digital ou *pen drive*;
- **Dilúvio dos dados:** Com as ciências computacionais, o volume de dados gerados cresce em uma escala nunca vista. Esta quantidade de dados é incompatível com a capacidade de análise e armazenamento disponíveis nas instituições (PHILIP CHEN; ZHANG, 2014);
- **Entropia dos dados:** O conhecimento sobre uma informação é gradativamente perdido após a conclusão do seu objetivo principal, como a publicação dos resultados de uma pesquisa (MICHENER et al., 1997). Com isso, sem a utilização de boas práticas para a gestão dos dados produzidos durante uma pesquisa, este conhecimento acaba sendo esquecido com o passar do tempo. A Figura 3 apresenta uma representação da entropia de dados científicos.

Figura 3 – Representação da entropia de dados científicos.

Fonte: Adaptado de (MICHENER et al., 1997).

- **Cauda longa (*long tail*) dos dados:** Em contraste ao volume de dados gerados pelos grandes projetos científicos, uma menor quantidade de dados, mas variada e distribuída, paralelamente é gerada por pesquisadores individuais e pequenos grupos de pesquisa. Diferente dos grandes projetos, que possuem várias equipes de pesquisadores, ferramentas computacionais modernas e procedimentos para gestão dos dados, estes pequenos projetos possuem recursos, ferramentas e pesquisadores limitados. Como consequência, os dados gerados nestes projetos geralmente não estão disponíveis na Web e são raramente reutilizados e preservados em longo prazo. Estas pequenos projetos formam o *Long Tail* da ciência (HEIDORN, 2008; HOWE et al., 2011; PALMER et al., 2007; WALLIS; ROLANDO; BORGMAN, 2013). A Figura 4 apresenta uma representação do *Long Tail* dos dados científicos. Conforme apresentado em Ferguson et al. (2014), os estudos demonstram uma distribuição distorcida entre o tamanho dos conjuntos de dados em relação a variedade de conjuntos de dados. Os grandes projetos científicos, com seus dados homogêneos e organizados representam apenas uma pequena proporção da variedade dos dados.

A grande proporção dos dados científicos está na *Long Tail*, com diversas iniciativas independentes de pesquisa, produzindo uma rica variedade de conjuntos de dados. Na extrema direita da *Long Tail* estão os dados não publicados ou indisponíveis. Apesar dos dados gerados pelos projetos que compõem o *Long Tail* representarem apenas uma pequena parte do volume dos dados científicos totais, estes dados formam a base de diversas teorias científicas (HEIDORN, 2008). Por isso, tornar os dados do *Long Tail* acessíveis e reutilizáveis é um dos grandes desafios da atualidade (WALLIS; ROLANDO; BORGMAN, 2013). Conforme descrito por Borgman (2012), o *Long Tail* também está tornando-se *data-intensive*, pois a geração de novos métodos e ferramentas adequados aos pequenos grupos de pesquisa, estão possibilitando uma otimização nos processos de coleta e geração de dados.

São necessárias melhorias que considerem todo o ciclo de vida dos dados, da captura e validação à curadoria, análise, compartilhamento e seu arquivamento permanente. Sem estes avanços, os dados continuarão sendo perdidos ou esquecidos, gerando perdas não somente científicas, mas também econômicas, devido a necessidade de repetição das coletas e dos experimentos anteriormente realizados. Em várias situações, como no caso de fenômenos ambientais, a perda dos dados é permanente, devido a impossibilidade de reprodução destes eventos (LYNCH, 2008).

Figura 4 - Esquema apresentando o *Long Tail* dos dados científicos.

Fonte: Adaptado de (FERGUSON et al., 2014).

2.2.4 Ciclo de Vida dos Dados

Para a gestão eficiente dos dados, é importante entender as diversas etapas pelo qual os dados passam durante o processo científico. Essa sequência de etapas é chamada do ciclo de vida dos dados. O uso deste ciclo é comum no gerenciamento de dados, pois auxilia na identificação do papel dos dados em cada uma das etapas de um processo, além de possibilitar a definição das melhores estratégias, ferramentas e boas práticas para a gestão dos dados ao longo do seu ciclo de vida (DAMA INTERNATIONAL, 2009; RÜEGG et al., 2014).

Com a evolução das ciências computacionais, a importância dos dados nas pesquisas foi intensificada, e o ciclo de vida tradicional dos dados científicos está sendo substituído por um novo ciclo de vida dos dados.

2.2.4.1 O Ciclo de Vida dos Dados Tradicional

Este ciclo de vida dos dados existe desde que a publicação de artigos científicos tornou-se um padrão há quase 400 anos (BRINEY, 2015). O ciclo inicia com o

planejamento da pesquisa, continua com a aquisição dos dados e sua análise, e termina com a publicação dos resultados da pesquisa. Este ciclo de vida é apresentado na Figura 5.

Figura 5 – Representação do antigo ciclo de vida dos dados.

Fonte: Adaptado de (BRINEY, 2015).

Neste processo, os dados possuem uma importante função nas etapas de aquisição e análise, que são atividades voltada aos dados. No entanto, os dados também estão envolvidos, com menor intensidade, nas etapas de planejamento do projeto e publicação. Por exemplo, no planejamento, é necessário a definição dos dados que serão coletados e, na publicação, os dados processados são apresentados em textos, figuras, gráficos e tabelas. Isso demonstra a importância dos dados em todas as etapas do processo de pesquisa tradicional, mas com maior intensidade nas etapas de aquisição e análise.

Este ciclo de vida mostra claramente que os dados são utilizados como um meio e não um fim. O objetivo do processo é a publicação de um artigo científico. Os dados são considerados subprodutos da pesquisa e geralmente são perdidos ou esquecidos após o término do projeto ou a publicação do artigo científico (BRINEY, 2015; VINES et al., 2014). Um estudo realizado por Vines et al. (2014) identificou que a disponibilidade dos dados após a publicação diminui 17% ao ano e que a probabilidade de encontrar um endereço de e-mail de trabalho dos autores de um estudo caem 7% ao ano. Portanto, quando mais antigo o artigo publicado, menos provável é a descoberta e o reuso dos seus dados.

Este é um dos maiores problemas do ciclo de vida dos dados atual, pois os dados de pesquisa, em geral, possuem um valor além da pesquisa original, tanto dentro do laboratório ou da instituição que realizou a pesquisa, quanto fora dela. No entanto, devido ao processo de pesquisa tradicional não valorizar o reuso dos dados, pouco

incentivo existe para a manutenção e a preservação destes dados após a sua publicação.

Apesar deste processo de trabalho ter funcionado durante centenas de anos, os avanços das ciências computacionais mostraram a importância dos dados e da sua preservação e reuso. Por isso, este cenário gerou uma necessidade de mudança do ciclo de vida dos dados científicos.

2.2.4.2 O Novo Ciclo de Vida dos Dados

Este ciclo de vida dos dados acrescenta etapas com foco no compartilhamento, na preservação e no reuso dos dados gerados no processo de pesquisa. Diversas instituições criaram seus próprios modelos de ciclo de vida dos dados, mas, em geral, seguem uma estrutura semelhante, voltada ao compartilhamento e reuso. Na Figura 6 é apresentado o modelo de ciclo de vida dos dados definido pelo DataONE.

Figura 6 – Modelo de ciclo de vida dos dados científicos do DataONE.

Fonte: Adaptado de (STRASSER et al., 2012).

Outro modelo de ciclo de vida dos dados, que demonstra de forma explícita o reuso dos dados, foi criado por Michener et al. (2011) e é apresentado na Figura 7. Neste modelo é visualizado a existência de dois ciclos. O primeiro ciclo inicia com a definição da pergunta de pesquisa, passando pela aquisição dos dados, a aplicação dos processos de qualidade, as análises e, finalmente, a publicação, que também produz novas perguntas de pesquisa, iniciando o ciclo novamente. O segundo ciclo surge a partir dos dados que foram preservados e compartilhados. Neste ciclo, novas perguntas de pesquisa são respondidas a partir do reuso dos dados existentes, que são integrados e analisados, gerando novas descobertas científicas, que também tem seus dados preservados. Este ciclo, baseado no reuso de dados, é mais ágil e menos custoso, impulsionando as novas descobertas científicas.

Figura 7 – Modelo de ciclo de vida dos dados científicos descrito em Michener et al. (2011).

Fonte: Adaptado de (MICHENER et al., 2011).

É possível concluir que a gestão de dados científicos exige esforços nas diversas etapas do ciclo de vida dos dados. Caso algumas destas etapas sejam negligenciadas, as etapas posteriores ficam incapacitadas de gerar resultados

eficientes. Portanto, para a geração de conhecimento científico reproduzível e reutilizável os esforços devem iniciar desde o planejamento da pesquisa, passando pelas diversas etapas intermediárias, até chegar na publicação, preservação e compartilhamento dos dados.

2.3 Metadados

A descrição dos dados, também chamada de documentação, é uma das etapas mais importantes do processo de gestão, pois possibilita a contextualização dos dados, necessária para que os mesmos possam ser localizados, compreendidos e consumidos por terceiros. Dados sem documentação se tornam dados sem significado e geralmente perdem seu valor. Apesar da existência de outros métodos tradicionais para descrição de dados em pesquisas, como os blocos de notas de papel, arquivos digitais, gravações de áudio ou vídeo, entre outros, o método de descrição considerado uma boa prática na gestão dos dados, e que cresce em popularidade nos últimos anos, é o uso dos metadados digitais (DAMA INTERNATIONAL, 2009).

A definição mais comum para metadados é que “metadados são dados sobre dados”. Conforme descrito em NISO (2004), metadados são informações estruturadas que descrevem, explicam, localizam e facilitam a recuperação, o uso e a gestão de um recurso de dados. Um recurso de dados é qualquer objeto que possa ser identificado, armazenado e descrito.

Os metadados podem descrever dados em diferentes níveis de abstração, como em um recurso simples, em parte de um recurso complexo ou em uma coleção de recursos. O nível ideal para aplicação dos metadados depende do tipo dos dados e da finalidade no qual os metadados serão empregados. Por exemplo, em um conjunto de dados científicos, os metadados podem ser aplicados no nível do agrupamento de registros, para descrever e contextualizar o conjunto de dados, mas também podem ser aplicados nos registros individualmente, para descrever informações específicas,

como a data/hora da geração do registro, onde foi gerado, por quem e em quais condições.

Os metadados são classificados em três tipos, conforme NISO (2004):

- **Metadados descritivos:** Descrevem os recursos para garantir sua descoberta e identificação. Estes metadados incluem elementos como título, resumo, autores e palavras-chave;
- **Metadados estruturais:** Indicam como os objetos de dados são organizados e interligados;
- **Metadados administrativos:** Disponibilizam informações que ajudam na gestão do recurso de dados. Estes metadados incluem informações sobre a origem do recurso, detalhes do seu formato, informações técnicas, entre outras.

Diferentes classificações de metadados são descritas na literatura, mas com várias semelhanças em comparação a apresentada por NISO (2004). Gilliland (2008) apresenta uma categorização com uma maior diversidade de tipos, no qual os metadados estruturais foram chamados de metadados técnicos e dois novos tipos (metadados de uso e metadados de preservação) foram derivados dos metadados administrativos, que permanecem presentes na categorização, mas com um foco mais restrito na descrição de informações sobre políticas, direitos de acesso e uso dos dados.

Os metadados podem ser embutidas no próprio recurso ou podem ser armazenados separadamente em outros arquivos ou bancos de dados. Por exemplo, em páginas HTML e em arquivos de imagens, os metadados geralmente são embutidos. No entanto, vários tipos de dados não permitem essa estratégia e precisam ser armazenados em arquivos auxiliares ou bases de dados.

Inúmeros padrões de metadados foram criados para diferentes tipos de dados científicos. Estes padrões são conhecidos como vocabulários.

2.3.1 Vocabulários

Os vocabulários, também chamados de esquemas de metadados ou padrões de metadados, são conjuntos de elementos estruturados para um objetivo específico (GREENBERG, 2005). O significado destes elementos é conhecido como a semântica do vocabulário. Os valores atribuídos a estes elementos são o seu conteúdo.

Um vocabulário define os elementos, também chamados de termos, utilizados de forma universal para determinado objetivo, permitindo que o compartilhamento destes dados se torne mais prático e seguro. Já as normas estabelecidas, permitem a definição de regras consistentes e vocabulários padronizados, que possibilitam uma comunicação eficiente entre os geradores e consumidores de dados. A estrutura padronizada também permite a pesquisa unificada destes dados, mesmo quando se encontram em bases de dados distribuídas.

As especificações destes vocabulários variam bastante, de diretrizes flexíveis, como o Dublin Core, para regras complexas e detalhadas, como o *Content Standard for Digital Geospatial Metadata* (CSDGM) do *Federal Geographic Documentation Committee* (FGDC) (GREENBERG, 2005). Geralmente estas especificações são direcionadas para um domínio ou uma aplicação específica, o que gera uma quantidade grande de vocabulários, dificultando a obtenção do seu principal objetivo, que é a padronização.

O vocabulário geralmente define nomes para os elementos presentes e um significado específico e bem definido. Essa definição é chamada de semântica do esquema (NISO, 2004). Adicionalmente são especificadas regras relacionadas ao conteúdo destes elementos, como o tipo e formato da informação aceito, sua cardinalidade, e a recomendação ou obrigatoriedade no uso de um vocabulário controlado.

2.3.2 Vocabulário Controlado

O vocabulário controlado é uma lista restrita de palavras ou termos utilizados para catalogação, marcação ou indexação. Ele possui este nome pois ao ser associado a um elemento de um vocabulário, restringe o conteúdo válido deste elemento aos termos presentes em uma lista restrita. Além disso, a edição de um vocabulário controlado só ocorre em condições especiais, a partir de um processo de revisão pela comunidade científica ou grupo regulador do vocabulário (HEDDEN, 2010).

Os vocabulários controlados são utilizados nos elementos de metadados descritivos para garantir a consistência, acurácia e uma eficiente indexação e recuperação de recursos digitais. Estes vocabulários são utilizados em sistemas de organização do conhecimento, como os vocabulários, thesaurus e ontologias.

A Tabela 1 apresenta o vocabulário controlado de tipos, definido pelo *Dublin Core Metadata Initiative* (DCMI). Este vocabulário controlado fornece uma lista de termos que podem ser utilizados como um conteúdo válido para o elemento *type*, que identifica a natureza ou o gênero do recurso descrito com o vocabulário *Dublin Core*.

Um vocabulário controlado pode ser facilmente incorporado em processos automatizados, simplificando atividades de entrada dos dados, a partir do fornecimento de listas selecionáveis para os usuários, além de suportar a validação de dados importados, provenientes de outros sistemas. Com o uso dos vocabulários controlados, a aplicação de técnicas para localização e análise dos dados tornam-se mais eficientes, devido a padronização e organização do seu conteúdo.

O processo de definição de um vocabulário controlado é realizado normalmente durante a criação de um vocabulário. Neste processo são definidos quais elementos do esquema necessitam de um vocabulário controlado. A partir dessa definição, os editores do vocabulário realizam as definições dos seus termos.

Tabela 1 - Vocabulário de tipos do DCMI, contendo o nome do termo, sua URI associada e uma definição do termo.

Termo	URI	Definição
Collection	<i>http://purl.org/dc/dcmitype/Collection</i>	Uma agregação de recursos.
Dataset	<i>http://purl.org/dc/dcmitype/Dataset</i>	Dados codificados em uma estrutura definida.
Event	<i>http://purl.org/dc/dcmitype/Event</i>	Uma ocorrência baseada no tempo.
Image	<i>http://purl.org/dc/dcmitype/Image</i>	Uma representação visual diferente de texto.
Interactive Resource	<i>http://purl.org/dc/dcmitype/Interactive Resource</i>	Um recurso que requer interação do usuário para ser compreendido, executado ou experimentado.
Moving Image	<i>http://purl.org/dc/dcmitype/Moving Image</i>	Uma série de representações visuais que transmite uma impressão de movimento, quando mostrado em sucessão.
Physical Object	<i>http://purl.org/dc/dcmitype/Physical Object</i>	Um objeto ou substância tridimensional inanimada.
Service	<i>http://purl.org/dc/dcmitype/Service</i>	Um sistema que fornece uma ou mais funções.
Software	<i>http://purl.org/dc/dcmitype/Software</i>	Um programa computacional na forma de código ou compilado.
Sound	<i>http://purl.org/dc/dcmitype/Sound</i>	Um recurso que pode ser ouvido.
StillImage	<i>http://purl.org/dc/dcmitype/StillImage</i>	Uma representação visual estática.
Text	<i>http://purl.org/dc/dcmitype/Text</i>	Um recurso que consiste principalmente de palavras para leitura.

Fonte: Adaptado de (DUBLIN CORE METADATA INITIATIVE, 2012).

Os vocabulários controlados podem ser de vários tipos (HEDDEN, 2010):

- **Lista de termos:** É a forma mais simples de vocabulário, no qual uma lista é definida com os termos permitidos. Exemplos de listas de termos são formatos de arquivos, linguagens, nomes de departamentos, etc.
- **Taxonomia:** É uma classificação hierárquica ou categorização de termos. É um vocabulário controlado no qual todos os termos pertencem a uma estrutura hierárquica e possuem relações de parentesco com outros termos. O sistema de classificação de Lineu, que classifica os seres vivos, é um exemplo de taxonomia (RAVEN; BERLIN; BREEDLOVE, 1971).
- **Thesaurus:** Também é uma classificação hierárquica. Cada termo possui relações com outros termos da estrutura, que podem definir conceitos de hierarquia, associação e equivalência entre estes termos.

2.3.3 Perfil de Aplicação

Um Perfil de Aplicação (PA), do inglês *Application Profile*, é um conjunto de elementos de metadados, regras e recomendações definidas para uma aplicação específica. Os elementos do PA podem ser de um ou mais esquemas de metadados, incluindo elementos definidos localmente, permitindo assim que uma determinada aplicação cumpra os seus requisitos funcionais (NISO, 2004; WOODLEY; CLEMENT; WINN, 2001).

Com o objetivo de suportar o processo de criação de PAs, a DCMI publicou a recomendação chamada de *Singapore Framework for Dublin Core Application Profiles* (NILSSON; BAKER; JOHNSTON, 2008), que define um conjunto de componentes descritivos que são necessários para a documentação de um PA, e descreve como esses documentos se relacionam. Os componentes do Perfil de Aplicação proposto foram baseados nesta recomendação.

2.3.3.1 Componentes do Perfil de Aplicação

De acordo com o *Singapore Framework*, o PA é uma documentação que consiste de um conjunto de componentes, com alguns sendo considerados obrigatórios e outros opcionais (NILSSON; BAKER; JOHNSTON, 2008):

- **Requisitos funcionais (obrigatório):** Os requisitos funcionais descrevem as funções que o PA foi projetado para suportar, assim como as funções fora do seu escopo. Os requisitos funcionais formam a base para a avaliação do PA quanto a sua consistência interna e para orientar sua correta utilização;
- **Modelo de Domínio (obrigatório):** O modelo de domínio define as entidades básicas descritas no PA e seus relacionamentos fundamentais. O propósito do modelo de domínio é definir o escopo básico de um PA. O modelo de domínio pode ser documentado em texto, diagramas ou usando uma abordagem formal, como o *Unified Modeling Language* (UML) (FOWLER, 2003);
- **Descrição do Perfil (*Description Set Profile – DSP*) (Obrigatório):** Define o conjunto de metadados que são instâncias válidas de um PA. O modelo DSP foi projetado para oferecer uma linguagem simples de restrições para os metadados do Dublin Core, com base no *DCMI Abstract Model (DCAM)* (POWELL et al., 2007). O DSP pode descrever restrições nos recursos, nas propriedades e na forma que um valor pode ser registrado.
- **Diretrizes de Uso (opcional):** Esta documentação descreve como utilizar o PA e como as propriedades devem ser consideradas no domínio de aplicação;
- **Diretrizes de Sintaxe da Codificação (opcional):** Esta documentação descreve regras e diretrizes sobre a codificação do PA.

Este modelo, e as relações de cada um dos componentes definido pelo *Singapore Framework*, é ilustrado na Figura 8.

Figura 8 – Modelo do DCAP segundo o *Singapore Framework*.

Fonte: Adaptado de (NILSSON; BAKER; JOHNSTON, 2008).

2.3.4 Sintaxe dos Metadados

Conforme descrito em Duval et al. (2002), a semântica é sobre o significado, enquanto a sintaxe é sobre a forma. Tanto a semântica quanto a sintaxe são essenciais para garantir o compartilhamento de dados e metadados entre dois sistemas. Mesmo com uma semântica bem definida, a falta de um alinhamento com a sintaxe torna o processo de compartilhamento complexo. No entanto, os esforços na área de metadados mostraram que a semântica e a sintaxe devem ser mantidas separadas. A rápida evolução de padrões da Web nos últimos anos demonstra essa necessidade.

Sintaxes comuns utilizadas para a codificação de metadados são os textos simples (*plain text*), *eXtensible Markup Language* (XML), *JavaScript Object Notation*

(JSON), *Resource Description Framework* (RDF) e bancos de dados. Em geral, cada uma destas sintaxes possui suas próprias ferramentas para criação e recuperação dos metadados.

No entanto, a escolha da sintaxe geralmente é definida pelo esquema de metadados, pois alguns esquemas foram criados baseados em padrões e tecnologias específicas.

2.3.5 Vocabulários na área de Biodiversidade

2.3.5.1 Darwin Core

O Darwin Core (DwC) (TDWG, 2013) é uma especificação de um vocabulário para dados biológicos. Este esquema foi definido para possibilitar a descrição da ocorrência de vida na Terra e suas relações com o meio ambiente, também chamados de dados primários de biodiversidade (WIECZOREK et al., 2012). Ele permite que ocorrências de espécies vivas na natureza e a existência de espécimes armazenadas em coleções biológicas (coleções físicas em museus ou arquivos digitais) possam ser representadas de forma padronizada.

A primeira versão do DwC foi proposta em 1999 e originalmente baseado no esquema de metadados *Dublin Core*. Por isso, é considerado uma extensão do *Dublin Core* para a área de biodiversidade. O DwC foi definido a partir de um conjunto de termos, que sofreu diversas revisões nos últimos anos, até transformar-se na versão atual do esquema, que se tornou um importante padrão internacional, utilizado por diversas instituições de todo mundo (WIECZOREK et al., 2012). Sua gestão é mantida pelo *Biodiversity Information Standards*, que anteriormente era chamado de *Taxonomic Scientific Working Group* (TDWG).

O DwC é baseado primariamente em táxons (que representa um grupo de indivíduos relacionados em um sistema de classificação de seres vivos) e tem como principal objetivo o suporte a descoberta, recuperação e integração destas informações entre sistemas e bancos de dados. O escopo do DwC está em constante

evolução, para possibilitar o seu uso em diferentes contextos, como listas taxonômicas e registros ambientais (ROBERTSON et al., 2014).

A popularidade deste vocabulário é atribuída também a existência de protocolos e ferramentas que suportam a sua utilização nos processos de compartilhamento e publicação dos dados. Destacam-se o *Darwin Core Archive* (DwC-A), um padrão para a estruturação e compartilhamento de registros DwC (BAKER; RYCROFT; SMITH, 2014), o *Integrated Publishing Toolkit* (IPT), uma ferramenta de publicação de arquivos DwC-A na Web (ROBERTSON et al., 2014), além das infraestruturas computacionais que consomem dados em DwC-A, como os portais de dados do GBIF (<http://www.gbif.org/>) e do *Atlas of Living Australia* (<http://www.ala.org.au>).

2.3.5.2 *Ecological Metadata Language*

O *Ecological Metadata Language*, (EML) (BLANKMAN; MCGANN, 2003; FEGRAUS et al., 2005), é uma especificação de um esquema de metadados para facilitar o compartilhamento, preservação e interpretação de dados científicos entre instituições e grupos de pesquisa. Ele é baseado nos trabalhos realizados pela *Ecological Society of America*, juntamente com parceiros de pesquisa, como o *Knowledge Network for Biocomplexity* (KNB). O EML foi estruturado de forma modular e extensível, para descrever tipos de dados provenientes das ciências ecológicas. Estes dados abrangem uma série de domínios distintos da ciência, incluindo estudos de biodiversidade, funções do ecossistema, meteorologia, ciências do solo, hidrologia, sensoriamento remoto e entre outras.

O EML permite que os pesquisadores forneçam descrições padronizadas de dados em um formato de texto acessível. As descrições do EML podem ser extremamente resumidas ou bastante ricas (incluindo descrições completas da sintaxe dos dados, esquema e semântica). Assim, o EML pode ser aplicado em uma grande variedade de aplicações de pesquisa, da descoberta de dados simples até o processamento automatizado dos dados.

2.3.5.3 Dublin Core

O *Dublin Core* (DC) (KUNZE; BAKER, 2007; WEIBEL et al., 1998) é um esquema de metadados para descrição de recursos digitais, tais como documentos, vídeos, sons, imagens, textos e sites da Web. Aplicações do DC consideram o XML e o RDF para a codificação dos seus metadados, apesar do seu uso em variadas sintaxes adicionais, como planilhas, arquivos texto e bancos de dados.

Este esquema de metadados é mantido pela DCMI, uma organização dedicada a promover a adoção de padrões de interoperabilidade e desenvolver vocabulários especializados para descrever recursos e tornar mais eficientes os sistemas de descoberta de dados.

O DC possui duas versões. O *Dublin Core Metadata Element Set*, mais antigo, que compreende um conjunto de 15 termos, e o *DCMI Metadata Terms*, que foi lançado com um conjunto de 55 termos, além de diversas mudanças nas regras e restrições do vocabulário.

Todos os seus elementos são opcionais e podem ser repetidos. Além disso, não existem restrições quanto a ordem dos elementos. Em sistemas de biodiversidade a utilização mais comum do DC é na descrição de documentos, artigos científicos e relatórios técnicos (SILVA et al., 2015).

2.3.5.4 Outros esquemas de metadados

Na Tabela 2, são apresentados outros vocabulários populares na comunidade de IB.

Tabela 2 - Lista de esquemas de metadados considerados em bases de dados de biodiversidade.

Esquema de Metadados	Objetivos	Referência
Plinian Core	Descrição de informações sobre espécies.	(BUITRAGO et al., 2016)
Audubon Core	Descrição de recursos e coleções multimídia de biodiversidade.	(MORRIS et al., 2013)
FGDC/CSDGM	Descrição de conjuntos de dados geoespaciais. Foi substituído pela ISO 19115 desde setembro/2010.	(METADATA AD HOC WORKING GROUP, 1998)
ISO 19115	Descrição de informações geográficas e serviços.	(ISO, 2003)
ABCD	Descrição de coleções biológicas.	(HOLETSCHEK et al., 2012)

Fonte: Autor.

2.3.6 Web Semântica

A partir da popularização dos vocabulários, vários sistemas computacionais e sítios na Web passaram a utilizar esta estratégia para descrever seus dados armazenados e publicados. Desta forma, desafios relacionadas a descrição padronizada de recursos e a interoperabilidade entre sistemas e bases de dados passaram a ser melhor gerenciadas.

No entanto, ainda era necessária uma base tecnológica que garantisse a interoperabilidade dos dados entre diferentes vocabulários e domínios de aplicação. A forma de estruturar e relacionar essa quantidade crescente de dados e possibilitar que não apenas as pessoas, mas as máquinas pudessem processar e raciocinar sobre estes dados tornou-se o principal desafio da Web Semântica. A partir das novas técnicas e padrões gerados no contexto da Web Semântica, soluções para estes desafios estão evoluindo nos últimos anos.

O conceito de Web Semântica surgiu em 2001, após a publicação de um artigo na revista *Scientific American*, intitulado “*The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*” (BERNERS-LEE et al., 2001). Neste artigo a Web Semântica foi definida como uma extensão da *World Wide Web*, na qual os dados possuem significados bem definidos, permitindo que pessoas e computadores possam trabalhar em cooperação.

A Web Semântica determina uma forma padronizada para expressar os relacionamentos entre páginas Web e dados, permitindo que não apenas humanos, mas também as máquinas, possam entender o significado das informações interligadas em um banco global de dados conectados.

Neste contexto, um conjunto de padrões foram especificados para possibilitar a identificação de recursos na Web, conforme apresentado na Figura 9.

Figura 9 – Camadas da Web Semântica.

Fonte: Adaptado de Nowack (2009).

Estas camadas da web semântica permitem a representação sintática, estrutural, semântica e lógica das informações referentes aos recursos armazenados. As camadas RDF e SPARQL serão descritas a seguir.

2.3.6.1 Resource Description Framework

O *Resource Description Framework* (RDF), criado pelo *World Wide Web Consortium* (W3C), é um modelo padronizado para representação e compartilhamento de recursos de dados na Web (CYGANIAK; WOOD; LANTHALER, 2014).

O RDF é formado por um conjunto de afirmações, do inglês *asserts*, sobre um recurso. Uma afirmação RDF consiste de três elementos (uma tripla) e tem a seguinte estrutura: sujeito, predicado e objeto, que visualmente é apresentada na Figura 10.

Figura 10 - Representação de uma afirmação em RDF, descrita por uma tripla.

Fonte: Autor.

Uma afirmação RDF expressa uma relação entre dois recursos, o sujeito e o objeto. O predicado representa a natureza da relação, que é definida de forma direcional (do sujeito para o objeto). O conjunto de triplas é chamado de grafo RDF, que pode ser representado visualmente por um grafo direcionado.

A identificação única e universal dos recursos e das propriedades descritas no RDF são realizadas por meio do *Uniform Resource Identifier* (URI), mais especificamente, o HTTP URI (BERNERS-LEE; FIELDING; MASINTER, 2005).

Os tipos mais comuns de URI são as *Uniform Resource Locators* (URLs, que possibilitam o acesso a uma representação visual do recurso descrito, como uma página Web, e os *Uniform Resource Names* (URNs), que são usados para identificar recursos, mas que não especificam uma localização ou forma de acesso ao recurso. Os URIs podem ser usados em qualquer uma das partes da tripla RDF (sujeito, objeto ou predicado).

Em uma tripla, o objeto também pode ter um valor textual. Nesta situação, a tripla é chamada de propriedade, pois representa uma propriedade do sujeito descrito.

A Figura 11 apresenta um grafo RDF, no qual a primeira tripla descreve o criador de um livro. O livro é identificado pela URI <http://example.com/book/1>, o autor é identificado pela URI <http://poli.usp.br/stu/dlins>, por fim, o predicado é representado pelo elemento *dcterms:creator*, do Dublin Core. A segunda tripla descreve o título do livro. Como o objeto não possui um URI, esta relação é uma propriedade do livro. Nesta tripla, o predicado é representado pelo elemento *dcterms:title*, do Dublin Core.

Figura 11 – Grafo RDF descrevendo um recurso.

Fonte: Autor.

Para dar um significado semântico para os predicados considerados nas triplas, são utilizados os vocabulários e ontologias, que definem descrições formais dos significados de cada elemento do esquema, diminuindo as ambiguidades da linguagem natural.

Estas descrições formais são identificadas por meio do *namespace* do vocabulário, representado também por uma URI. Por exemplo, o *Dublin Core Terms* é representado pelo *namespace* <http://purl.org/dc/terms/> e pelo prefixo “dcterms”. O prefixo pode ser utilizado para simplificar a representação de um *namespace*. Na Figura 11 é possível visualizar a utilização do *namespace* do *Dublin Core Terms* para descrever o autor e o título de um livro. O predicado identificado como *dcterms:creator* também pode ser representado por <http://purl.org/dc/terms/creator>.

Múltiplos *namespaces* podem ser utilizados na descrição de um mesmo recurso, possibilitando que elementos de diferentes vocabulários possam ser combinados em uma mesma descrição. Múltiplas descrições, criadas em momentos diferentes, com propósitos distintos, também podem ser ligadas umas às outras por meio do RDF.

O RDF pode ser codificado ou serializado por diferentes representações sintáticas, algumas mais adequadas para o processamento de máquina, e outras mais legíveis para as pessoas. Entre as notações mais populares estão o RDF/XML (GANDON; SCHREIBER; BECKETT, 2014), o Turtle (BECKETT et al., 2014), o N-Triples (BECKETT, 2014) e o JSON-LD (SPORNY et al., 2014). Na Figura 12 é apresentado um grafo RDF serializado em Turtle, uma das representações do RDF mais legíveis e populares.

Figura 12 - Grafo RDF em Turtle, referente as triplas apresentadas na Figura 11.

```
@prefix dcterms: <http://purl.org/dc/terms/> .  
  
<http://example.com/book/1> dcterms:creator  
<http://poli.usp.br/stu/dlins> .  
<http://example.com/book/1> dcterms:title "Data Provenance" .
```

Fonte: Autor.

2.3.6.2 SPARQL

Os dados na Web Semântica utilizam o modelo do RDF para sua representação e podem ser armazenados em bancos de dados de triplas, em bancos de dados relacionais, que possuam um mecanismo de mapeamento para RDF, e em arquivos RDF publicados na Web.

O SPARQL, acrônimo de *SPARQL Protocol and RDF Query Language*, é um conjunto de especificações que fornece linguagens e protocolos para consulta e manipulação do conteúdo de grafos RDF na Web ou em banco de dados (DUCHARME, 2013). Em 2008, o SPARQL 1.0 tornou-se uma recomendação oficial da W3C. E em 2013, o SPARQL 1.1 substitui a recomendação anterior (THE W3C SPARQL WORKING GROUP, 2013).

Na Figura 13 é demonstrado a utilização do SPARQL para a consulta de um grafo RDF. As variáveis SPARQL começam com um caractere “?” e podem ser definidas em qualquer elemento da tripla (sujeito, predicado ou objeto). Tanto na cláusula SELECT quanto na cláusula WHERE é possível definir triplas, juntamente com as variáveis, para a encadeamento das regras da consulta. Além disso, o

SPARQL possibilita a inclusão de filtros e operadores para possibilitar a realização de consultas complexas.

Figura 13 – Demonstração do uso do SPARQL para a realização de uma consulta em um grafo RDF.

Grafo RDF

```
@prefix dcterms: <http://purl.org/dc/terms/> .  
@prefix foaf: <http://xmlns.com/foaf/0.1/> .  
  
<http://example.com/book/1> dcterms:creator <http://poli.usp.br/stu/dlins> .  
<http://example.com/book/1> dcterms:title "Data Provenance" .  
<http://poli.usp.br/stu/dlins> foaf:name "Daniel Lins" .
```

Consulta

```
PREFIX dcterms: <http://purl.org/dc/terms/>  
PREFIX foaf: <http://xmlns.com/foaf/0.1/>  
SELECT ?nome  
WHERE  
{  
  ?livro dcterms:title "Data Provenance" .  
  ?livro dcterms:creator ?autor .  
  ?autor foaf:name ?nome.  
}
```

Resultado

nome
"Daniel Lins"

Fonte: Autor.

Bancos de dados de triplas e frameworks que manipulam dados em RDF oferecem pontos de acesso via Web, compatíveis com o protocolo SPARQL e sua linguagem de consulta. Esses pontos de acesso são denominados SPARQL *endpoints*. Um SPARQL *endpoint* aceita consultas SPARQL e retorna os resultados via protocolo *Hypertext Transfer Protocol* (HTTP) (PRUD'HOMMEAUX; BUIL-ARANDA, 2013).

2.4 Proveniência dos Dados

2.4.1 Conceitos Gerais

O termo proveniência foi inicialmente utilizado no contexto da arte para descrever a história documentada ou a cadeia de custódia de uma obra de arte, da sua criação até o seu mais recente proprietário (JAFFE, 1996; MOREAU et al., 2008; TAN, 2007). A proveniência permite que os especialistas possam comprovar a autenticidade de um objeto e influencia diretamente no seu valor (MOREAU et al., 2008). Por isso, muito do trabalho de determinação da proveniência de uma obra de arte consiste na localização, análise e julgamento da autenticidade da documentação existente. Esta tarefa envolve um complexo trabalho de análise do objeto, juntamente com a pesquisa em catálogos, fotos de arquivos, correspondências e registros históricos. Adicionalmente, este trabalho torna-se mais complexa devido aos falsos registros gerados, por exemplo, quando obras de arte são roubadas, contrabandeadas ou ilegalmente descobertas. No entanto, caso a proveniência completa de uma obra seja conhecida, este é um indicador de qualidade e valorização do objeto (PANCERELLA; MYERS; RAHN, 2002).

Três características gerais da proveniência podem ser identificadas a partir do seu uso na arte:

- A proveniência é demonstrada por uma documentação precisa e contínua;
- A proveniência considera dados do passado;
- A proveniência trata da rastreabilidade de processos;
- A proveniência é caracterizada pela confiabilidade e a integridade dos dados.

Na arte, os processos relevantes para a proveniência são as movimentações da obra de arte durante a sua cadeia de custódia, entre proprietários e casas de leilão, e durante atividades de restauração. E a documentação deste processo é composta geralmente por documentos em papel, que acompanham a obra de arte.

A proveniência também é aplicada nas mais variadas áreas, como nas bibliotecas digitais, na indústria de alimentos, no jornalismo, na rastreabilidade de informações em redes sociais e na transparência de aplicações comerciais (AKABANE; LOPES; SILVA, 2011; CURBERA et al., 2008; MOREAU; GROTH, 2013).

Na proveniência de dados científicos, os processos considerados são bastante variados. Mas assim como na arte, conhecer os processos que geram um objeto digital garante uma maior confiabilidade e permite um melhor entendimento dos dados e dos processos que o gerou. No entanto, na Web ou em sistemas de grande escala, multiusuários e com rápido fluxo das informações, a proveniência documentada em papel não é viável. Por isso, métodos computacionais estão sendo propostos para possibilitar a construção de uma proveniência qualificada, de modo que ela possa ser recuperada, analisada e bem fundamentada (MOREAU et al., 2011).

Diferentes sinônimos são empregados para o termo proveniência de dados, como *data lineage*, *pedigree*, *data tracking* e *data provenance*, sempre referindo-se ao histórico dos dados e a descrição do processo de produção do mesmo (LIU; ÖZSU, 2009). Várias definições de proveniência também foram apresentadas na comunidade científica. Buneman et al. (2000) afirma que proveniência é a documentação complementar de um determinado dado que contém a descrição de “como”, “quando”, “onde” e “por que” o dado foi obtido e “quem” o obteve. Segundo Simmhan et al. (2005), a proveniência dos dados está relacionado aos processos de investigação e armazenamento da origem do dado, bem como sua movimentação entre os bancos de dados. Por fim, Belhajjame et al. (2013) afirma que a proveniência é o registro que descreve pessoas, instituições, entidades e atividades envolvidas na produção, influência ou entrega de um dado ou objeto. Esta última definição é a considerada mais adequada no contexto deste trabalho.

2.4.2 Modelos de proveniência dos dados

O conceito de proveniência foi investigado sob variados nomes pela comunidade da ciência da computação desde os anos oitenta. Na atualidade, um recente

direcionamento das pesquisas de proveniência foi na sua representação e compartilhamento, de modo a descrever a origem de recursos de dados na Web (MOREAU et al., 2015). Isto resultou em eventos da comunidade científica para buscar um entendimento sobre a essência da proveniência e a definição de um modelo de dados padronizado. Destes esforços foi criado um modelo de proveniência chamado *Open Provenance Model* (OPM) (MOREAU et al., 2011).

Posteriormente, a partir dos esforços do Grupo de Trabalho em Proveniência (*Provenance Working Group*) do W3C, em 2013 foi publicado o PROV, um modelo de dados para proveniência na Web, considerado a evolução do OPM (KWASNIKOWSKA; MOREAU; BUSSCHE, 2015). Apesar do OPM ter sido utilizado em várias iniciativas da comunidade científica, a partir da publicação da PROV como recomendação da W3C, o PROV passou a ser utilizado em grande parte das novas iniciativas da comunidade científica.

Neste trabalho, o PROV será considerado para a descrição da proveniência dos dados de experimentos científicos. O modelo do PROV será detalhado na próxima seção.

2.4.2.1 Modelo W3C PROV

O PROV foi definido pela W3C como um modelo para a descrição e troca de dados de proveniência na Web. Este trabalho resultou na geração e publicação de treze documentos, incluindo quatro documentos normativos, seis documentos técnicos, um relatório de implementação e dois documentos para descrição do padrão (GROTH; MOREAU, 2013), (MISSIER; BELHAJJAME; CHENEY, 2013). A família de documentos do PROV descreve um conjunto de especificações que permitem que a proveniência seja modelada, serializada, compartilhada, acessada, integrada, traduzida e fundamentada.

A ontologia PROV-O (*PROV Ontology*) expressa a estrutura do PROV por meio da *Web Ontology Language* (OWL2) (SMITH et al., 2012). Ela fornece um conjunto de classes, propriedades e restrições que podem ser utilizadas para representar e

compartilhar informações de proveniência geradas em diferentes sistemas e em diferentes contextos (LEBO et al., 2013). O PROV também pode ser organizado em um grupo de conceitos fundamentais, que possuem os elementos que formam a essência da proveniência, e um conjunto de construções estendidas com subtipos e relações expandidas, que podem ser usadas em aplicações mais complexas. O PROV também permite a sua extensão a partir da criação de novas classes e propriedades para satisfazer requisitos específicos de aplicações (MISSIER; BELHAJJAME; CHENEY, 2013).

Os conceitos fundamentais do PROV são apresentados na Tabela 3 e descritos abaixo:

- **Entity (Entidade):** Entidade é um objeto que se deseja descrever. Pode ser físico, digital ou qualquer outro tipo de objeto. Exemplos: registros, conjuntos de registros, artigos científicos, arquivos ou modelos;
- **Activity (Atividades):** Atividade é algo que ocorre em um determinado período de tempo consumindo e/ou transformando entidades. Pode ser descrita pela duração e sua correlação com as entidades envolvidas. Geralmente as atividades são de consumo, processamento, transformação, modificação, transferência, utilização e geração de entidades;
- **Agents (Agentes):** Agente é algo ou alguém que assume alguma forma de responsabilidade no processo descrito. Pode ser responsável pelas atividades realizadas, por entidades existentes ou pelas atividades realizadas por outros agentes. Exemplos: cientistas, instituições, grupos de pesquisa;
- **Relações:** Relações são as propriedades que descrevem afirmações entre as classes (tipos) do PROV (entidades, atividades e agentes).

Tabela 3 – Mapeamento dos conceitos fundamentais do PROV.

Conceitos	Tipos e Relações	Elementos do PROV
Entidade	Classes do PROV	<i>Entity</i>
Atividade		<i>Activity</i>
Agente		<i>Agent</i>
Geração	Relações do PROV	<i>WasGeneratedBy</i>
Uso		<i>Used</i>
Comunicação		<i>WasInformedBy</i>
Derivação		<i>WasDerivedFrom</i>
Atribuição		<i>WasAttributedTo</i>
Associação		<i>WasAssociatedWith</i>
Delegação		<i>ActedOnBehalfOf</i>

Fonte: Adaptado de Belhajjame et al. (2013).

O diagrama da Figura 14, representa as três classes do PROV e as propriedades que as relacionam. As construções estendidas do PROV são descritas de forma detalhada em Belhajjame et al. (2013) e Lebo et al. (2013).

Como uma especificação para a proveniência, o PROV foi criado com o objetivo de ser aplicado em diversos cenários de uso. Cada um destes cenários possui perspectivas diferentes sobre a proveniência e também diferentes necessidades de informações que precisam ser capturadas e descritas nos seus registros.

Figura 14 – As três classes principais do PROV e suas relações.

Fonte: Lebo et al. (2013).

Conforme descrito por Belhajjame et al. (2013), as três perspectivas principais consideradas no PROV são:

- **Visão de Responsabilidade:** A proveniência está preocupada com a descrição dos autores que realizaram ou participaram de uma atividade ou processo;
- **Visão do Fluxo do Processo:** A proveniência considera a descrição do fluxo, ou sequência, das atividades dentro de um processo;
- **Visão do Fluxo de Dados:** A proveniência considera a descrição dos recursos de dados, suas transformações e derivações.

Estas perspectivas podem ser combinadas com base nas necessidades do cenário no qual o PROV é aplicado. A Figura 15 apresenta um diagrama publicado em da Silva et al. (2016) para representar a descrição de um processo de monitoramento ambiental utilizando o modelo de dados do PROV.

Conforme pode ser verificado, mesmo um simples processo de coleta e publicação de dados pode gerar informações valiosas, essenciais para a avaliação e

reuso dos dados publicados. Algumas das perguntas que podem ser respondidas a partir desta proveniência são:

- Como foi realizada a coleta destes dados? De forma manual ou automatizada? Quais equipamentos foram considerados?
- Quem foram os pesquisadores envolvidos neste processo?
- Os dados publicados são brutos ou passaram por transformações? Quais são os detalhes da atividade de transformação aplicada aos dados brutos?
- Onde posso acessar os dados brutos coletados?

Figura 15 – Descrição em alto nível de um processo de monitoramento ambiental, por meio dos elementos fundamentais do modelo de dados do PROV.

Fonte: da Silva et al. (2016).

Este exemplo demonstra o valor que a publicação de metadados de proveniência podem gerar aos dados publicados. Neste trabalho, o PROV será considerado, juntamente com vocabulários adicionais reconhecidos pela comunidade científica,

para a criação de um perfil de aplicação voltado para a gestão da proveniência de dados na área da biodiversidade.

2.4.2.2 Outras iniciativas de proveniência e extensões do PROV

Além dos padrões PROV e OPM, descritos anteriormente, iniciativas anteriores também foram criadas com o objetivo de estruturar e padronizar os dados de proveniência de processos. Entre eles destacam-se o Provenir (SAHOO; SHETH, 2009), o *Provenance Web* (HARTIG, 2009) e o PASOA (GROTH; MILES; MOREAU, 2009).

Outras iniciativas também apresentam extensões para os modelos de dados existentes, com o objetivo de satisfazer requisitos de um domínio de aplicação específico. O D-PROV (MISSIER et al., 2013) e o PROV-Wf (COSTA et al., 2013) são extensões do PROV que possibilitam a descrição da proveniência prospectiva, com foco na sua utilização em ferramentas de gerenciamento de workflow científico (WfMS). O PAV, sigla de *Provenance, Authoring and Versioning*, (CICCARESE et al., 2013) é uma ontologia que estende o PROV com o objetivo de descrever com mais detalhes informações sobre autoria, curadoria e a criação digital de recursos na Web. O SEIS-PROV, é uma extensão do PROV para a área de sismologia.

2.4.3 Soluções computacionais para a Proveniência de Dados

Uma arquitetura para gestão da proveniência de dados contém três componentes principais: Um mecanismo de captura, o modelo de dados de proveniência e uma infraestrutura para armazenar, descobrir e consultar os dados (FREIRE et al., 2008).

Os mecanismos de captura de proveniência variam dos manuais até os totalmente automatizados e podem ser divididos em 3 grupos: sistemas baseados em *workflow*, sistemas baseado no sistema operacional e sistemas baseados em processos (FREIRE et al., 2008).

Os sistemas baseados em workflow são os mais difundidos nas pesquisas científicas. Estes componentes são conectados a um Sistema Gerenciador de Workflow, do inglês *Workflow Management System* (WfMS). Estudos propuseram padrões, técnicas e ferramentas para aprimorar a captura dos dados de proveniência nos WfMS e permitir a avaliação e a reprodutibilidade dos experimentos (ALTINTAS; BARNEY; JAEGER-FRANK, 2006; CALLAHAN et al., 2006; WOLSTENCROFT et al., 2013).

Destacam-se o Vistrails (CALLAHAN et al., 2006; FREIRE et al., 2014), o Kepler (ALTINTAS; BARNEY; JAEGER-FRANK, 2006; DOU et al., 2012) e o Taverna (MISSIER et al., 2010; WOLSTENCROFT et al., 2013).

As vantagens destes mecanismos são a rápida preparação da ferramenta para a captura dos metadados de proveniência e a possibilidade de reprodução automática dos experimentos e processos executados no WfMS. Como estes mecanismos de proveniência utilizam as mensagens do próprio motor do WfMS, para registrar a proveniência, a captura dos dados é realizada de forma bastante transparente para o pesquisador.

A principal desvantagem desta abordagem é a dependência de ferramentas, pois estes mecanismos são criados para WfMS específicos. Outra desvantagem é a ausência de interoperabilidade entre os mecanismos de proveniência de cada WfMS e também de outros tipos de ferramentas, dificultando a manutenção dos metadados de proveniência em ambientes distribuídos e heterogêneos. Por fim, a maioria dos sistemas baseados em workflow armazenam seus metadados de proveniência em uma base de dados ou arquivo local, o que dificulta o compartilhamento destes dados com outros pesquisadores.

Os sistemas baseados no sistema operacional não exigem esforços para preparação ou alterações dos scripts e ferramentas utilizadas. A coleta da proveniência é realizada no nível do sistema operacional, durante a execução do processo. Neste grupo destacam-se o PASS (MUNISWAMY-REDDY et al., 2006) e o ES3 (FREW; SLAUGHTER, 2008).

Embora estes mecanismos possam capturar os dados de forma transparente e sem esforços adicionais de configuração e anotação, esta abordagem gera informações de baixo nível e em grande volume, que necessita ser processada para poder gerar informações viáveis para análise (FREIRE et al., 2008).

Por fim, os sistemas baseados em processos, necessitam que os metadados sejam documentados de forma manual, por meio de anotações. Estes mecanismos permitem a seleção dos metadados de proveniência que serão capturados e a definição da granularidade ideal dos metadados, conforme as necessidades do domínio de aplicação. Neste grupo destacam-se Sumatra (DAVISON et al., 2014), o NoWorkflow (MURTA et al., 2014; PIMENTEL et al., 2015), o YesWorkflow (MCPHILLIPS et al., 2015), o Preserv (GROTH; MILES; MOREAU, 2005).

As desvantagens destes mecanismos são os esforços adicionais necessários para a preparação e documentação dos metadados. Em geral, estes mecanismos são acompanhados de ferramentas e *plug-ins* de captura, que diminuem esta carga de trabalho do pesquisador.

Em relação a infraestrutura para armazenar, descobrir e consultar os metadados de proveniência, diversas abordagens foram empregadas em estudos e iniciativas da comunidade científica. Vários mecanismos de proveniência baseados em workflow também possuem recursos para a consulta e visualização dos dados. Destacam-se o Vistrails e o Kepler.

o ProvStore (HUYNH; MOREAU, 2014) é um repositório de proveniência que suporta o padrão PROV para a representação da proveniência de dados e atividades. Ele utiliza uma abordagem baseada em um banco de dados de documentos, no qual documentos PROV podem ser armazenados, acessados e visualizados na Web ou por meio de uma Web API.

Outra iniciativa interessante é o MyExperiment (GOBLE et al., 2010), um ambiente de pesquisa online para o compartilhamento de *workflows*. Os *workflows* podem ser gerados no Taverna, Kepler, entre outras ferramentas. Apesar do MyExperiment não armazenar a proveniência retrospectiva, que é gerada a partir da

execução do experimento, ela armazena a proveniência prospectiva, gerada durante o desenho do *workflow*.

A arquitetura proposta nesta tese permite o uso dos vários tipos de mecanismos de captura apresentados nesta seção, devido a estratégia que define um componente desacoplado apenas para a descrição e aquisição dos metadados de proveniência. O processamento e armazenamento dos metadados de proveniência são realizados por meio dos serviços web do servidor de proveniência, que podem ser acessados por WfMS, *scripts* e aplicações variadas. No caso de aplicações que não possuem acesso direto aos serviços web, *plug-ins* e componentes adicionais podem ser implementados, para permitir essa comunicação e garantir a padronização dos metadados de proveniência. Relacionado ao armazenamento, disponibilização e visualização dos metadados, a arquitetura proposta armazena os metadados em um banco de triplas, no qual os metadados de proveniência são organizados em grafos. Também são disponibilizados serviços para a consulta da proveniência de experimentos e de recursos de dados, que permitem que os pesquisadores acessem estas informações pela Web em formatos variados. O detalhamento da arquitetura computacional proposta será apresentado no Capítulo 4.

3 PERFIL DE APLICAÇÃO PARA PROVENIÊNCIA DE DADOS DE BIODIVERSIDADE

A comunidade de IB utiliza vocabulários para a padronização dos dados e interoperabilidade entre seus sistemas e bases de dados. No entanto, estas iniciativas não consideram, entre seus principais objetivos, a descrição de metadados de proveniência.

Em contrapartida, o PROV e os demais vocabulários e ontologias que descrevem a proveniência de dados, são genéricos e foram criados para a representação da proveniência de dados em variados domínios e aplicações.

Neste cenário, a estratégia considerada neste trabalho foi a definição de um PA gerado a partir do reuso de conceitos descritos por ontologias e vocabulários existentes e reconhecidos pela comunidade científica. Este PA contempla os elementos necessários para satisfazer os requisitos da arquitetura computacional proposta neste trabalho, além de possibilitar níveis de interoperabilidade entre soluções que consideram estes vocabulários.

Apesar desta premissa, foi encontrada uma grande dificuldade para a reutilização de vocabulários publicados em artigos científicos. Os principais motivos verificados foram:

- Ausência de documentação suficiente para o uso do vocabulário;
- Ausência de *namespace* válido e/ou das definições formais do vocabulário em RDF/OWL, de forma aberta, na Web.

Em Malta e Baptista (2014), que apresenta um estudo sobre os PAs criados na última década pela comunidade científica, estes problemas também são citados, confirmando esta dificuldade como um dos desafios na geração de perfis de aplicação.

Com isso, durante a construção do PA, chamado de *Biodiversity Provenance* (BioProv), foram considerados os conceitos definidos na ontologia do PROV, para a descrição das relações entre recursos de dados, processos, pessoas e instituições,

envolvidas no ciclo de vida dos dados. Adicionalmente, foram incluídas propriedades de outros vocabulários existentes, como o *Dublin Core*, o *Darwin Core* e o FOAF, para a descrição dos objetos descritos, garantindo a contextualização dos dados e processos. Para os conceitos no qual não foram identificados vocabulários e propriedades possíveis de reutilização, foram definidos elementos locais no PA. Além disso, foram definidos vocabulários controlados para determinar o correto preenchimento de algumas propriedades do PA, facilitando a compreensão destes metadados e sua posterior consulta.

A versão atual deste PA não espera cobrir totalmente todas as técnicas e processos de análise e experimentação existentes na comunidade de IB. Um perfil de aplicação necessita de tempo para sua evolução e amadurecimento. A partir de novos estudos de caso e validações com a comunidade científica de IB espera-se que este PA possa continuar evoluindo, em busca de um modelo ideal para o registro da proveniência de dados nesta área.

3.1 Processo de Desenvolvimento do Perfil de Aplicação para Proveniência de Dados de Biodiversidade

O processo utilizado no desenvolvimento do PA foi baseado no Me4DCAP, apresentada em Malta & Baptista (2013). Neste processo são definidas as atividades necessárias para a geração dos componentes (*Deliverables*) descritos no *Singapore Framework*.

3.1.1 Especificação dos Requisitos Funcionais

O objetivo do PA para proveniência de dados de biodiversidade é a definição de um conjunto de metadados que possam descrever, de forma objetiva, os experimentos e demais processos envolvidos na geração de recursos de dados neste domínio de aplicação.

Não é objetivo deste PA ser um substituto dos vocabulários atualmente utilizados pela comunidade de IB para descrição dos dados, como o Darwin Core e o EML. Este PA busca a complementação dos metadados disponibilizados atualmente, que possuem uma lacuna em relação a descrição dos processos que envolvem a geração de um recurso de dados.

A definição do escopo e dos requisitos do PA proposto iniciou com o detalhamento do processo de experimentação científica, com o objetivo de identificar as informações essenciais para a garantia da proveniência destes processos. A Figura 16 apresenta um esquema representando os elementos essenciais de um experimento científico, com base nos requisitos identificados.

Figura 16 – Esquema representando as informações necessárias para descrever um experimento científico.

Fonte: Autor.

Com base no esquema apresentado na Figura 16, foram identificados os elementos, ou classes, que agrupam as informações dos experimentos científicos:

- **Experimento:** Representa as informações relacionadas ao experimento propriamente dito. Com estas informações é possível a contextualização do experimento. Delimitação do problema, objetivos, hipóteses, etc.;

- **Projeto de Pesquisa:** Representa as informações relacionadas ao projeto de pesquisa no qual o experimento foi realizado. Nem todo experimento necessita estar associado a um projeto de pesquisa, e um projeto de pesquisa pode ter um ou mais experimentos associados a ele;
- **Atividade:** Representa as informações relacionadas as atividades envolvidas no processo. As informações sobre as atividades são subdivididas em elementos como dados, responsáveis, transformações, ambiente de execução, etc.;
- **Resultado/Produto:** Representa as informações sobre os produtos e resultados gerados no experimento e no projeto de pesquisa, como publicações, conjuntos de dados, modelos, entre outros.

Como requisitos funcionais fundamentais para construção do PA foram considerados:

- Registrar todos os processos de transformação de um recurso de dados, incluindo informações sobre insumos utilizados (dados de entrada), ferramentas, algoritmos, parâmetros de configuração e o ambiente de execução, além dos dados de saída (produtos gerados);
- Registrar informações sobre o contexto do experimento, como decisões e premissas consideradas durante a sua execução. Estas informações geralmente são negligenciadas no processo de documentação de um experimento, mas são valiosas para a avaliação e entendimento do processo por terceiros;
- Registrar metadados sobre o experimento propriamente dito. Qual seus objetivos, hipóteses consideradas, equipe de participantes, se o mesmo está associado a um projeto de pesquisa, entre outras informações relacionadas ao experimento;
- Registrar os responsáveis pelas diversas atividades de um experimento, possibilitando a rastreabilidade de responsabilidades. Além disso, identificar os responsáveis pela descrição dos metadados. Desta forma,

gerando credibilidade para as metadados de proveniência disponibilizadas.

3.1.2 Desenvolvimento do Modelo de Domínio

Após a definição dos requisitos funcionais do PA para proveniência dos dados de biodiversidade, foi criado um modelo de domínio. O Modelo de domínio é a descrição de quais elementos o PA irá descrever e os relacionamentos entre estes elementos (BAKER; COYLE, 2009). O Modelo de domínio do PA proposto é apresentado na Figura 17.

Figura 17 – Modelo de domínio do Perfil de Aplicação para proveniência de dados de Biodiversidade.

Fonte: Autor.

Conforme visualizado na Figura 17, os principais elementos dos experimentos científicos foram incluídos no PA como classes. A partir daí, foram definidas as relações entre cada uma delas e suas propriedades.

Um **experimento** pode pertencer a um **projeto de pesquisa**. O **experimento** contém **atividades**. Os **responsáveis** executam as **atividades**, mas elas também podem ser iniciadas por meio de outras **atividades**. Uma **atividade** usa **recursos**, que podem ser dados, equipamentos, sistemas, algoritmos ou qualquer outro insumo para a realização de uma atividade. A **atividade** também gera **recursos**, que podem ser modelos, publicações, dados atualizados, mapas, etc. Um **recurso** pode possuir outros **recursos**. Por exemplo, um conjunto de dados possui registros. O **recurso** também pode ser associado a um **responsável**. Por fim, um **responsável** pode representar outro **responsável**. Este caso é aplicado na relação entre pesquisador e instituição, no qual o vínculo do pesquisador com a instituição existe durante a realização da atividade.

Após a definição desta primeira versão do modelo, foi realizada a sua adequação ao modelo de dados do PROV. Com isso, foi possível a geração de um modelo compatível ao PROV, garantindo assim, a sua interoperabilidade com outras aplicações baseadas neste vocabulário. A Figura 18 apresenta o novo modelo de domínio, considerando as classes e relações do PROV.

Figura 18 – Modelo de domínio do Perfil de Aplicação para proveniência de dados de Biodiversidade.

Fonte: Autor.

Neste novo modelo, foram criadas as subclasses *bioprov:DataResource* e *bioprov:Resource*, ambas herdadas da classe do PROV *prov:Entity*. A primeira representa os recursos de dados e a segunda representa os demais recursos (algoritmos, ferramentas, etc.) considerados nas atividades de um experimento científico. A partir da *prov:Entity* também foi criada a subclasse *bioprov:Project*, que representa o projeto de pesquisa no qual o experimento está relacionado. A classe *bioprov:Experiment*, é uma subclasse da classe *prov:Bundle*, que têm como função principal agrupar um conjunto de metadados de proveniência. Por fim, as classes *bioprov:Researcher* e a *bioprov:Organization* foram criadas como subclasses da *prov:Agent*, para representar de forma mais específica, os pesquisadores e as organizações envolvidos nos processos descritos.

3.1.3 Desenvolvimento do *Description Set Profile* (DSP)

Após o refinamento do modelo de domínio, que considerou uma especialização do PROV, foram analisadas, em outros vocabulários, possíveis propriedades que pudessem enriquecer as informações coletadas para cada uma das classes definidas no modelo.

Em relação aos experimentos científicos e projetos de pesquisa, diversas ontologias e vocabulários foram analisadas, sendo identificadas boas iniciativas para a descrição deste tipo de dados. Destacaram-se a *Ontology of Scientific Experiments* (EXPO) (SOLDATOVA; KING, 2006), a *Core Scientific Concepts* (CoreCS) (DUMA et al., 2016), a *Open Provenance Ontology* (OvO) (DA CRUZ; CAMPOS; MATTOSO, 2012) e o *Ecological Metadata Language* (EML) (FEGRAUS et al., 2005).

Vocabulários como o *Dublin Core*, o *Darwin Core*, o *Friend of a Friend* (FOAF) (BRICKLEY; MILLER, 2014), e o *RDF Schema* (BRICKLEY; GUHA; MCBRIDE, 2014), também tiveram propriedades utilizadas para a descrição das classes do PA. Todos os vocabulários utilizados no PA são apresentados na Tabela 4.

Tabela 4 – Vocabulários considerados no PA proposto, com os respectivos *namespaces* e prefixos.

Vocabulário	Namespace	Prefixo
PROV-O	http://www.w3.org/ns/prov#	prov
Dublin Core Terms	http://purl.org/dc/terms/	dcterms
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	rdf
RDF Schema	http://www.w3.org/2000/01/rdf-schema#	rdfs
Darwin Core	http://rs.tdwg.org/dwc/terms/	dwc
Dublin Core	http://purl.org/dc/elements/1.1/	dc
Friend of a Friend	http://xmlns.com/foaf/0.1/	foaf
Biodiversity Provenance	http://bioprov.poli.usp.br/	bioprov
Geonames	http://www.w3.org/2003/01/geo/wgs84_pos#	geo
Time Ontology	http://www.w3.org/2006/time#	time
OWL	http://www.w3.org/2002/07/owl#	owl

Fonte: O Autor.

Após a definição dos vocabulários a serem considerados no PA, o DSP foi detalhado para descrever as suas classes, relações e propriedades. A Tabela 5 apresenta a entidade de experimento científico descrita no DSP.

Tabela 5 – Descrição da entidade Experimento Científico no DSP do perfil de aplicação BioProv.

Entidade	Experimento Científico	Classe	<i>Bioprov:Experiment</i>
Descrição	Propriedade	Abrangência	Textual
identificador	<i>dcterms:identifier</i>	-	sim
título	<i>dcterms:title</i>	-	sim
descrição	<i>dcterms:description</i>	-	sim
tipo da entidade	<i>dcterms:type</i>	-	sim
justificativa	<i>bioprov:cause</i>	-	sim
delimitação do Problema	<i>bioprov:problemDelimitation</i>	-	sim
hipóteses	<i>bioprov:hypothesis</i>	-	sim
objetivos	<i>bioprov:goals</i>	-	sim
metodologia	<i>bioprov:method</i>	-	sim
requisitos	<i>bioprov:has_requirements</i>	-	sim
cobertura taxonômica	<i>bioprov:taxonomicCoverage</i>	-	sim
cobertura temporal	<i>dcterms:temporal</i>	-	sim
cobertura espacial	<i>dcterms:spatial</i>	-	sim
data/hora de início	<i>prov:startedAtTime</i>	<i>xsd:dateTime</i>	não
data/hora de término	<i>prov:endedAtTime</i>	<i>xsd:dateTime</i>	não
responsável	<i>prov:wasAttributedTo</i>	<i>prov:Agent</i>	não
Resultados	<i>bioprov:has_results</i>	-	sim

Fonte: O Autor.

A versão completa do DSP, considerada no momento da publicação deste trabalho, é apresentada no Apêndice A.

4 ARQUITETURA COMPUTACIONAL PARA GESTÃO DE METADADOS DE PROVENIÊNCIA

A definição de uma arquitetura computacional tem por objetivo evidenciar a estrutura necessária para a compreensão de um sistema (BASS; CLEMENTS; KAZMAN, 2012). Este capítulo descreve uma arquitetura computacional com o objetivo de subsidiar futuras implementações de um sistema distribuído para gestão de metadados de proveniência, compreendendo os elementos de software necessários e as relações entre eles, assim com suas respectivas características.

A descrição da arquitetura computacional foi baseada no Modelo de Referência para Processamento Distribuído Aberto, do inglês *Reference Model of Open Distributed Processing* (RM-ODP). Este modelo considera a especificação de uma arquitetura a partir de cinco pontos de vista, no qual se analisa e constrói a arquitetura: empresarial, informação, computacional, engenharia e tecnologia (ISO/IEC 10746-1, 1998; ISO/IEC 10746-2, 2009; ISO/IEC 10746-3, 2009; ISO/IEC 10746-4, 1998).

A Figura 19 apresenta os pontos de vista do RM-ODP e os aspectos da arquitetura considerados em cada um deles:

- **Ponto de vista empresarial:** Descreve as necessidades dos usuários do sistema, trazendo uma visão sobre as funções do sistema dentro do cenário que ele estará aplicado;
- **Ponto de vista da informação:** Representa os elementos de informação considerados no sistema;
- **Ponto de vista da computação:** Descreve a visão computacional, considerando os componentes da arquitetura e como os mesmos estão organizados, suas interfaces e relações;
- **Ponto de vista da engenharia:** Descreve as estruturas de gerenciamento e comunicação, que garantem o suporte as aplicações distribuídas;
- **Ponto de vista da tecnologia:** Descreve os componentes (hardware e software) que compõem o sistema distribuído. Neste ponto de vista

também é realizado o mapeamento dos pontos de vista de computação e engenharia em tecnologias específicas.

Figura 19 - Pontos de Vista do RM-ODP.

Fonte: Adaptado de (JEBBAR; SEKKAKI; BENAMAR, 2012).

4.1 Requisitos para a Proveniência dos Dados

Nesta seção são apresentados os requisitos (funcionais e não funcionais) consideradas na construção da arquitetura computacional proposta. Estes requisitos foram identificados com base na literatura científica, em discussões com pesquisadores da área de IB e a partir das avaliações dos resultados dos experimentos realizados durante a pesquisa.

Os requisitos identificados foram organizados nas três dimensões da proveniência de dados definidos em Groth (2012): conteúdo, gerenciamento e uso. Cada requisito é representado por um código com prefixo e numeração sequencial. Sendo o prefixo RF para requisito funcional e o prefixo RN para requisito não-funcional.

4.1.1 Requisitos de Conteúdo

Estes requisitos apresentam as necessidades relacionadas aos metadados de proveniência. Ou seja, quais as informações precisam ser armazenadas para garantir que as informações essenciais sobre a origem dos dados estarão disponíveis para posterior consulta e análise.

4.1.1.1 RF01 - Identificar unicamente os recursos registrados nos metadados de proveniência

A arquitetura deve possibilitar que o usuário associe identificadores aos objetos registrados no repositório de proveniência (recursos, atividades e responsáveis), para possibilitar que os mesmos possam ser identificados unicamente, recuperados e reutilizados de forma eficiente, além de evitar problemas relacionados a duplicação de registros.

Caso o usuário não possua este identificador único para atribuir ao recurso descrito, a arquitetura deve realizar a geração de um identificador único e associá-lo ao recurso, informando para o usuário o identificador atribuído.

4.1.1.2 RF02 – Armazenar metadados descritivos juntamente com os metadados de proveniência

A arquitetura deve possibilitar que o usuário registre metadados descritivos, juntamente com os metadados de proveniência. Os metadados descritivos são importantes para a contextualização dos dados e dos processos registrados, além de facilitar o entendimento e a consulta aos metadados. Preferencialmente estes elementos de metadados devem fazer parte de padrões de metadados reconhecidos pela comunidade científica, facilitando assim sua correta compreensão e utilização.

4.1.1.3 RF03 – Armazenar a identificação do usuário responsável pela geração dos metadados de proveniência

A arquitetura deve armazenar, em um log ou base de dados, o usuário responsável pelo registro dos metadados no repositório de proveniência. Os metadados que descrevem a ação de registro da proveniência, e também chamados de proveniência da proveniência, são informações essenciais para a credibilidade destes metadados, pois armazenam informações sobre quem, quando e como estes registros de proveniência foram gerados.

4.1.1.4 RF04 – Armazenar metadados sobre autoria e licença de uso dos recursos

A arquitetura deve armazenar metadados sobre a autoria e a licença de uso de um recurso. A primeira, para possibilitar o reconhecimento dos envolvidos na geração e evolução dos recursos de dados. A segunda, para garantir a disponibilização das regras para reutilização destes dados por terceiros.

4.1.1.5 RN01 - Definir um modelo de dados padronizado

A arquitetura deve considerar um modelo de dados padronizado para a descrição da proveniência dos dados. A partir da utilização de um modelo padronizado, a interoperabilidade com outras ferramentas será facilitada, assim como a análise e a comparação de recursos e processos.

4.1.1.6 RN02 - Possibilitar a descrição dos metadados de proveniência em variadas granularidades

A arquitetura deve possibilitar que o usuário descreva a proveniência de dados em diferentes níveis, conforme o objetivo da proveniência. Por exemplo, em algumas situações a proveniência pode ser descrita no nível do conjunto de dados, mas em outras, as descrições são necessárias no nível de registros.

4.1.1.7 RN03 – Tornar os registros de proveniência permanentes e imutáveis

A arquitetura deve considerar que os metadados de proveniência são imutáveis enquanto estiverem disponíveis no repositório de proveniência. No entanto, a exclusão de registros deve ser permitida a partir da definição de um prazo para expiração dos metadados de proveniência de um recurso.

Os dados sem prazo de expiração devem ser mantidos permanentemente no repositório de proveniência.

4.1.2 Requisitos de Gerenciamento

Estes requisitos contemplam as questões relacionadas aos mecanismos de captura, gerenciamento e compartilhamento dos metadados de proveniência na Web.

4.1.2.1 RF05 – Apoiar a reprodutibilidade e o reuso

A arquitetura deve apoiar a reprodutibilidade de um experimento a partir da manutenção dos dados e metadados registrados durante a execução deste processo. A gestão destes metadados e seu compartilhamento de forma aberta possibilita que estes dados e processos possam ser avaliados, reproduzidos e reutilizados.

4.1.2.2 RF06 – Possibilitar consultas distribuídas em repositórios de proveniência

A arquitetura deve realizar consultas distribuídas ou federadas, isto é, que consideram a junção de metadados armazenados em diferentes repositórios de proveniência. Desta forma, possibilitando que metadados de um repositório possam ser complementados com informações disponibilizadas em outros repositórios. Por exemplo, instituições de pesquisa e universidades podem gerenciar repositórios de proveniência internos para seus experimentos. No entanto, estes repositórios podem

compartilhar seus metadados para que consultas federadas possam construir toda a cadeia de reutilização de um recurso de dados.

4.1.2.3 RF07 – Armazenar arquivos descritos nos metadados de proveniência

A arquitetura deve possibilitar o armazenamento de arquivos consumidos e gerados durante um processo. Este armazenamento pode ser realizado no servidor de proveniência ou a partir da integração com serviços de armazenamento na nuvem. Com isso, garantindo que todos os recursos descritos nos metadados de proveniência, independentemente de serem insumos, produtos intermediários ou produtos finais do experimento, possam ser localizáveis e reutilizáveis.

4.1.2.4 RF08 – Consultar a proveniência de um recurso publicado na Web

A arquitetura deve disponibilizar serviços Web para que os metadados de proveniência de um recurso de dados possam ser recuperados e consumidos por pesquisadores e outros sistemas computacionais.

4.1.2.5 RF09 – Identificar automaticamente a proveniência de um recurso visualizado na Web

A arquitetura deve possibilitar que durante a consulta de um recurso de dados, publicado na Web, o usuário possa identificar a existência de metadados de proveniência sobre este recurso e decidir se deseja visualizar estas informações.

4.1.2.6 RF10 – Autenticar e monitorar os usuários que registram metadados de proveniência

A arquitetura deve exigir a autenticação dos usuários responsáveis pelo registro de metadados de proveniência. Desta forma, possibilitando o registro do usuário e de suas ações (4.1.1.3 RF03).

4.1.2.7 RF11 – Exportar os metadados de proveniência

A arquitetura deve possibilitar a exportação dos metadados de proveniência de um recurso, para que possam ser consumidos por outras aplicações.

4.1.2.8 RF12 – Importar metadados de proveniência gerados por outras aplicações

A arquitetura deve possibilitar a importação de metadados de proveniência gerados por outras aplicações de proveniência que utilizem padrões compatíveis aos considerados na arquitetura descrita.

4.1.2.9 RN04 – Possibilitar o uso da proveniência em ambientes heterogêneos

A arquitetura deve ser capaz de gerenciar metadados de proveniência descritos e registrados por diferentes ferramentas. Por isso, a arquitetura não deve ser baseada em ferramentas específicas, o que dificultaria sua aplicação em ambientes heterogêneos, como a comunidade de IB.

4.1.2.10 RN05 – Considerar protocolos e padrões da Web

A arquitetura deve ser baseada em protocolos e padrões internacionalmente reconhecidos para a publicação e compartilhamento de dados na Web.

4.1.3 Requisitos de Uso

Estes requisitos contemplam as necessidades relacionadas ao uso dos metadados de proveniência.

4.1.3.1 RF13 – Possibilitar consultas nos metadados de proveniência

A arquitetura deve possibilitar que o usuário execute consultas no repositório de proveniência, garantindo que estes metadados possam ser recuperadas e analisados com base nas necessidades dos usuários.

4.1.3.2 RF14 – Consultar a proveniência completa de um recurso

A arquitetura deve possibilitar a consulta e visualização de todos os metadados de proveniência de um recurso, a partir do seu identificador único. A proveniência completa de um recurso contempla os metadados registrados durante o todo processo de geração ou utilização do recurso.

4.1.3.3 RF15 – Consultar os metadados de proveniência de um experimento

A arquitetura deve possibilitar a consulta e visualização de todos os metadados relacionados a um experimento, a partir do seu identificador único. A proveniência de um experimento contempla os metadados registrados durante todas as etapas de um experimento, incluindo os recursos utilizados e gerados, suas atividades e agentes envolvidos.

4.1.3.4 RF16 – Consultar e recuperar produtos intermediários gerados e armazenados durante a execução de um experimento.

A arquitetura deve possibilitar a consulta e recuperação dos produtos de dados intermediários gerados e armazenados no repositório de proveniência durante o experimento.

4.1.3.5 RF17 – Consultar a proveniência das derivações de um recurso

A arquitetura deve possibilitar a consulta da proveniência das derivações de um recurso (*forward provenance*), isto é, a descrição de como um recurso foi utilizado desde a sua criação ou quais os novos recursos gerados a partir deste recurso.

4.1.3.6 RF18 – Consultar os softwares, algoritmos e parâmetros utilizados em um experimento científico

A arquitetura deve possibilitar a consulta e visualização de informações detalhadas sobre um experimento científico, contemplando os softwares e algoritmos utilizados, parâmetros e dados de entrada. Estas informações são essenciais para a reprodutibilidade de um experimento ou a validação de um recurso de dados, a partir da auditoria do seu processo de geração.

4.2 Pontos de Vista do RM-ODP

Baseado nos requisitos apresentados nas seções anteriores, a especificação da arquitetura computacional para a gestão dos metadados de proveniência foi construída. Nesta seção, a arquitetura computacional será apresentada seguindo os cinco pontos de vista definidos pelo RM-ODP.

4.2.1 RM-ODP: Ponto de Vista Empresarial

Este ponto de vista apresenta as justificativas e objetivos da arquitetura computacional proposta baseada nas necessidades dos seus usuários. No contexto deste trabalho, a empresa corresponde a comunidade internacional de IB, envolvida nos processos de geração, gestão e análise de dados de biodiversidade. A justificativa e os objetivos para esta arquitetura computacional foram apresentados e amplamente discutidos ao longo deste trabalho. Estudos envolvendo aspectos desta problemática foram publicados em Silva et al. (2016) e Silva, Batista e Corrêa (2016).

4.2.2 RM-ODP: Ponto de Vista de Informação

Este ponto de vista apresenta quais informações são gerenciadas pela arquitetura computacional e como elas são organizadas e manipuladas. No contexto deste trabalho, o modelo de dados considerado é baseado no perfil de aplicação BioProv, descrito no Capítulo 3.

4.2.3 RM-ODP: Ponto de Vista Computacional

O ponto de vista computacional apresenta os componentes da arquitetura, sua organização e responsabilidades. A arquitetura proposta, apresentada na Figura 20, é composta por quatro componentes principais:

- Ferramentas de aquisição e consulta;
- Servidor de proveniência;
- Repositório de metadados;
- Repositório de arquivos.

A **ferramenta de aquisição** realiza a descrição ou captura dos metadados de proveniência relacionados aos experimentos científicos. Esta ferramenta pode ser representada por uma biblioteca em uma linguagem de programação, um sistema computacional, uma ferramenta de workflow ou um *plug-in*.

Após a descrição ou captura dos metadados de proveniência, a ferramenta envia estes metadados para o **servidor de proveniência** por meio dos **serviços Web** disponíveis. Além dos metadados, esta ferramenta também envia arquivos associados aos recursos descritos na proveniência.

O **servidor de proveniência**, ao receber os metadados realiza sua validação, processamento e armazenamento no **repositório de metadados**. Os arquivos enviados são armazenados no **repositório de arquivos**. Por fim, a **ferramenta de consulta** realiza a requisição de relatórios de proveniência sobre recursos de dados e experimentos. Estas ferramentas também podem submeter consultas

personalizadas ao **servidor de proveniência** e requisitar a recuperação de arquivos relacionados aos recursos de dados.

Figura 20 – Representação dos componentes da arquitetura computacional e suas relações.

Fonte: Autor.

A organização dos serviços de proveniência foi definida com base no perfil de aplicação BioProv, no qual as classes do PA representam os *endpoints* dos serviços. Além disso, serviços adicionais são disponibilizados para a geração dos relatórios de proveniência, para as consultas SPARQL, para o armazenamento de arquivos e para a importação de metadados de proveniência gerados por outras ferramentas. A

Tabela 6 apresenta os *endpoints* dos serviços, suas descrições e os métodos HTTP disponíveis.

Tabela 6 – Descrição dos *endpoints* dos serviços de proveniência da arquitetura proposta.

<i>Endpoints dos serviços</i>	<i>Descrição</i>	<i>Métodos</i>
/experiments/	Registra, consulta e exclui metadados de experimentos científicos. O experimento é um agregador de metadados de proveniência.	GET POST DEL
/projects/	Registra, consulta e exclui projetos de pesquisa que podem agrupar um ou mais experimentos científicos.	GET POST DEL
/resources/	Registra, consulta e exclui recursos de dados e demais recursos consumidos e gerados em um experimento científico.	GET POST DEL
/activities/	Registra, consulta e exclui atividades. As atividades são eventos que ocorrem durante o experimento científico, que utilizam e/ou geram recursos.	GET POST DEL
/agents/	Registra, consulta e exclui agentes. Os Agentes são responsáveis pelos recursos e pela execução de atividades. Pode representar uma pessoa ou instituição.	GET POST DEL
/provenance/<uri>	Consulta a proveniência complete de um recurso (relatório de proveniência).	GET
/experiments/<uri>/data/	Consulta todos os metadados de proveniência relacionados a um experimento.	GET
/sparql?query=<query>	SPARQL <i>Endpoint</i> . Consulta o repositório de metadados.	GET
/entity/<uri>/files/	Registra, consulta e exclui arquivos associados a recursos registrados no repositório de proveniência.	GET POST DEL
/rdf/	Registra arquivos RDF/XML, permitindo a importação de conjuntos de metadados de proveniência gerados por outras ferramentas ou sistemas legados.	POST

/users/	Cadastra, consulta e excluir usuários para acesso aos serviços de proveniência.	GET POST DEL
/token/	Geração de token único a partir da autenticação de um usuário cadastrado. Depois de gerado este token pode ser utilizado pelo usuário a cada nova requisição aos serviços de proveniência	GET

Fonte: O Autor.

O servidor de proveniência também possui componentes adicionais que realizam as demais funções necessárias para o funcionamento da arquitetura computacional:

- **Gestor:** Módulo principal do sistema que gerencia o funcionamento dos demais componentes do servidor de proveniência.
- **Segurança:** Autenticação e monitoramento dos usuários que registram metadados no repositório de proveniência. Estes usuários são cadastrados em uma base de dados exclusiva deste componente.
- **Modelo de Dados:** Validação e transformação dos metadados armazenados, com base no perfil de aplicação BioProv;
- **Gerador de URI:** Geração de URIs para a identificação única e global dos recursos descritos no repositório de proveniência;
- **Conversor e Serializador:** Conversões entre os formatos considerados na arquitetura computacional. Os metadados em JSON-LD, provenientes dos serviços, são convertidos para o formato RDF/XML, para serem armazenados no repositório de metadados. A conversão contrária é realizada quando consultas ao repositório de metadados são executadas.

Os metadados de proveniência, após serem recebidos pelos serviços Web, validados e convertidos para o formato RDF/XML, são armazenados em um banco de dados de triplas (*triple store*), localizado no repositório de metadados. A Figura 21 apresenta um grafo RDF simples, que corresponde a proveniência de um processo de atualização de um conjunto de dados (*Dataset*).

Figura 21 – Exemplo de grafo RDF representado graficamente e no formato RDF/XML.

Fonte: Autor.

Todos os recursos de dados descritos na arquitetura de proveniência necessitam estar associados a um identificador único global, representado por um HTTP URI.

Para garantir o registro da proveniência de um experimento, sem exigir que o pesquisador possua previamente estes identificadores definidos, a arquitetura disponibiliza um componente para geração de identificadores únicos. Com isso, ao

descrever os metadados de um recurso, o pesquisador pode definir o identificador do recurso ou delegar a geração do identificador para a arquitetura computacional.

Vários padrões universais de identificação, utilizados pela comunidade científica, são compatíveis com os HTTP URI e podem ser utilizados na arquitetura para a identificação dos recursos, como o DOI e o *Archival Resource Key* (ARK) (KUNZE, 2003).

4.2.4 RM-ODP: Ponto de Vista de Engenharia

O ponto de vista de engenharia apresenta como são definidas as interações entre os componentes distribuídos da arquitetura computacional.

Os quatro principais componentes da arquitetura computacional, descritos na seção 4.2.3, podem ser dispostos em um ambiente distribuído. A comunicação entre estes componentes é realizada por meio de serviços compatíveis com o estilo arquitetural *Representational State Transfer* (REST) (FIELDING, 2000).

Este estilo arquitetural é representado por uma comunicação cliente/servidor, no qual suas requisições e respostas são baseadas no envio de uma representação de um recurso. Na arquitetura proposta, todas as comunicações utilizam o HTTP, mas diferem no formato de codificação dos dados. Os serviços de proveniência utilizam o JSON-LD e o RDF/XML. Os serviços do repositório de arquivos, utilizam o JSON. Por fim, o repositório de metadados utiliza o RDF/XML e o SPARQL (THE W3C SPARQL WORKING GROUP, 2013). A arquitetura proposta disponibiliza também um *SPARQL Endpoint*, que permite que consultas no protocolo SPARQL sejam realizadas no repositório de proveniência. A partir do *SPARQL Endpoint*, também é possível a execução de pesquisas federadas, que realizam consultas simultâneas em repositórios, com a consolidação dos seus resultados. A Figura 22 apresenta a distribuição dos componentes da arquitetura, seus protocolos e formatos de comunicação.

Figura 22 – Representação da distribuição dos componentes da arquitetura computacional proposta.

Fonte: Autor.

Os serviços de proveniência utilizam o formato JSON-LD para a serialização dos metadados. Apenas o serviço de importação de documentos RDF que utiliza o formato RDF/XML. Esta decisão considerou as seguintes questões:

- O JSON-LD permite a serialização de triplas e grafos RDF;
- O JSON-LD é 100% compatível com o JSON, que é um dos formatos mais populares na Web. Com isso, o uso do JSON-LD possibilita que os sistemas atuais, mesmo que incompatíveis com RDF, possam consumir os serviços de proveniência;
- O uso do JSON-LD facilita o desenvolvimento de *plug-ins* e ferramentas adicionais para a arquitetura proposta, pois as principais tecnologias são compatíveis e possuem ferramentas para manipular documentos JSON;
- O número de profissionais habilitados a desenvolver soluções considerando serviços no estilo arquitetural REST e documentos JSON é maior que os desenvolvedores habilitados nas tecnologias da Web Semântica (LANTHALER; GÜTL, 2012).

Para que ferramentas e sistemas possam consumir os serviços de proveniência de forma simples, outras estratégias no uso do formato JSON-LD foram consideradas.

O conceito de contexto, definido na especificação do formato (SPORNY et al., 2014), possibilita que a definição semântica dos metadados sejam descritas em um documento externo (LANTHALER; GÜTL, 2012). Tornando o documento JSON-LD semelhante a um documento JSON padrão, mas com a especificação do objeto *@context*, que referencia o documento externo que possui a descrição semântica do documento.

Na Figura 23 é apresentado um documento JSON-LD que descreve a proveniência de um conjunto de dados. Neste documento, a URI do conjunto de dados, suas propriedades e relações são descritas a partir da utilização de termos simples (*id*, *title*, *subject*, *identifier*, *created*, *wasDerivedFrom*). A semântica destes metadados, é descrito a partir de um documento de contexto, publicado na Web no endereço *http://bioprov.poli.usp.br/prov.jsonld*, e referenciada com o *@context*.

Figura 23 – Documento JSON-LD descrevendo a proveniência de um conjunto de dados.

```
{
  "@context": "http://bioprov.poli.usp.br/prov.jsonld",
  "id": "http://bioprov.poli.usp.br/dataset/2",
  "title": "Birds Dataset",
  "subject": "Dataset of species occurrences of birds of north Brazil",
  "identifier": "occ_ds_birds_amazon",
  "created": "2014-02-25",
  "wasDerivedFrom": "http://bioprov.poli.usp.br/dataset/1"
}
```

Fonte: Autor.

O documento de contexto utilizado na arquitetura computacional descreve semanticamente todos o perfil de aplicação BioProv e associa seus elementos a termos simples, facilitando a descrição dos metadados pelas ferramentas que irão consumir os serviços de proveniência.

A Figura 24 apresenta uma versão simplificada do documento de contexto utilizado na arquitetura computacional de proveniência.

Figura 24 – Documento de contexto, que descreve semanticamente os termos utilizados em um documento JSON-LD.

```
{
  "@context": {
    "bioprov": "http://bioprov.poli.usp.br/",
    "dc": "http://purl.org/dc/elements/1.1/",
    "dcterms": "http://purl.org/dc/terms/",
    "prov": "http://www.w3.org/ns/prov#",
    "title": "dcterms:title",
    "id": { "@type": "@id", "@id": "@id" },
    "subject": "dcterms:subject",
    "created": "dcterms:created",
    "identifier": "dcterms:identifier",
    "agent": { "@type": "@id", "@id": "prov:agent" },
    "entity": { "@type": "@id", "@id": "prov:entity" },
    "activity": { "@type": "@id", "@id": "prov:activity" },
    "wasGeneratedBy": { "@type": "@id", "@id": "prov:wasGeneratedBy" },
    "generated": { "@reverse": "prov:wasGeneratedBy", "@type": "@id" },
    "generation": { "@type": "@id", "@id": "prov:qualifiedGeneration" },
    "wasInformedBy": { "@type": "@id", "@id": "prov:wasInformedBy" },
    "wasDerivedFrom": { "@type": "@id", "@id": "prov:wasDerivedFrom" },
    "wasRevisionOf": { "@type": "@id", "@id": "prov:wasRevisionOf" }
  }
}
```

Fonte: Autor.

No documento de contexto, cada definição representa uma associação entre um elemento de um vocabulário com um termo simples. Por exemplo, a definição *"title": "dcterms:title"* significa que o termo *title* representa o elemento *http://purl.org/dc/elements/1.1/title* do Dublin Core.

Já no caso da definição *"wasDerivedFrom": { "@type": "@id", "@id": "prov:wasDerivedFrom" }*, significa que o termo *wasDerivedFrom* representa o elemento *http://www.w3.org/ns/prov#wasDerivedFrom* do PROV e que este elemento deve ser interpretado como uma URI.

Ao receber os metadados submetidos pelos serviços de proveniência, o servidor utiliza o documento de contexto para realizar a validação e conversão destes metadados em um grafo RDF, descrevendo a semântica destes metadados. Além disso, por meio do serviço requisitado, o servidor de proveniência identifica a classe do recurso a ser descrito. Por exemplo, ao registrar a proveniência utilizando o *endpoint /resources/*, o servidor de proveniência irá processar este recurso como uma

entidade (*prov:Entity*) e associar os metadados descritos no documento JSON-LD como propriedades e relações deste recurso.

A Figura 25 apresenta o documento de proveniência da Figura 23 no formato RDF/XML, após sua validação e processamento pelo servidor de proveniência. Neste documento é possível identificar a classe do recurso, suas propriedades e sua relação de derivação. Após este processo o grafo RDF é armazenado no repositório de metadados da arquitetura computacional.

Figura 25 – Metadados após sua conversão pelo servidor de proveniência.

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:prov="http://www.w3.org/ns/prov#"
  xmlns:dc="http://purl.org/dc/terms/">

  <prov:Entity rdf:about="http://bioprov.poli.usp.br/dataset/1">
    <dc:title>Birds Dataset</dc:title>
    <dc:created>2014-02-25</dc:created>
    <prov:wasDerivedFrom
rdf:resource="http://bioprov.poli.usp.br/dataset/2"/>
    <dc:subject>Dataset of species occurrences of birds of north
Brazil</dc:subject>
    <dc:identifier>occ_ds_birds_amazon</dc:identifier>
  </prov:Entity>

</rdf:RDF>
```

Fonte: Autor.

4.2.5 RM-ODP: Ponto de Vista de Tecnologia

O ponto de vista de tecnologia deve focar no desenvolvimento do sistema, identificando as ferramentas e os padrões utilizados para a sua implementação. Nesta seção é apresentado o protótipo funcional desenvolvido para a realização dos estudos de caso e validação da estratégia computacional proposta. Os estudos de caso são descritos no Capítulo 5.

A Figura 26 apresenta os componentes da arquitetura computacional que foram implementados no protótipo, juntamente com as tecnologias utilizadas.

Figura 26 – Representação dos componentes do protótipo funcional da arquitetura computacional de proveniência.

Fonte: Autor.

O protótipo funcional desenvolvido é composto por três aplicações e um serviço na nuvem que representam os quatro componentes principais da arquitetura computacional proposta.

O servidor de proveniência, é uma aplicação Web que disponibiliza uma API REST com os serviços para registro e consulta dos metadados de proveniência. Esta aplicação foi desenvolvida em Python (<https://www.python.org/>), juntamente com o *microframework* Flask (<http://flask.pocoo.org/>). Os componentes internos do servidor

de proveniência foram desenvolvidos como classes Python e realizam as funções essenciais determinadas pela arquitetura computacional. Também foram utilizados diversos bibliotecas Python para o apoio ao desenvolvimento da aplicação, com destaque para a RDFLib (<https://rdflib.readthedocs.io/>), utilizada na manipulação, validação e conversão de documentos RDF;

As demais classes de negócio da aplicação foram organizadas pelos elementos do modelo de proveniência: experimentos, entidades, atividades, agentes, arquivos, usuários e proveniência. Para o armazenamento do cadastro de usuários da aplicação foi utilizado um banco de dados relacional sqlite (<https://www.sqlite.org/>).

O acesso ao repositório de metadados, representado pelo *triple store* AllegroGraph (<http://franz.com/agraph/allegrograph/>), é realizado por meio de uma REST API disponibilizada pela própria ferramenta.

Para o armazenamento de arquivos associados aos recursos descritos no experimento foi utilizado o servidor de armazenamento S3 da Amazon (<https://aws.amazon.com/pt/s3/>).

Para a descrição dos metadados de proveniência durante experimentos científicos foi desenvolvido um pacote na linguagem R (<https://www.r-project.org/>), que disponibiliza funções que podem ser utilizadas em qualquer projeto R para a descrição e armazenamento de metadados de proveniência, gerados durante a execução do projeto.

Este pacote R utiliza os serviços Web do servidor de proveniência para o armazenamento dos metadados de proveniência descritos. Além das funções de descrição, o pacote disponibiliza funções para a consulta de metadados e a geração de relatórios de proveniência.

Por fim, foi desenvolvido uma aplicação Javascript (<https://www.javascript.com/>) como extensão (*plug-in*) do navegador Google Chrome, que permite a visualização da proveniência de recursos publicados em páginas Web. Esta aplicação analisa automaticamente os metadados (incluídos nas *tags* HTML) das páginas abertas no

navegador e, ao localizar um metadado de proveniência, utiliza os serviços de proveniência para consultar os metadados descritos do recurso publicado.

Os repositórios com os códigos-fonte das aplicações desenvolvidas estão disponíveis no Apêndice B.

5 VALIDAÇÃO DA PROPOSTA

Este capítulo apresenta e discute os resultados de dois estudos de caso utilizados na validação do perfil de aplicação BioProv e da arquitetura computacional para gestão de metadados de proveniência. O tipo de experimento escolhido para os estudos de caso foi a modelagem de distribuição de espécies (MDE).

A MDE possui uma grande importância na área da biodiversidade, com aplicações na predição da distribuição de espécies raras ou ameaçadas de extinção, na proposta de reintrodução de espécies, no planejamento de áreas prioritárias para conservação, na análise de efeitos das mudanças climáticas sobre a biodiversidade, entre outras aplicações (CORRÊA et al., 2011; DE MARCO JÚNIOR; SIQUEIRA, 2009).

A MDE pode utilizar exclusivamente dados ambientais (dados abióticos), que definem as características do *habitat*, tais como as condições climáticas, tipo de solo e altitude, como também dados adicionais, como as interações entre espécies (dados bióticos) e os movimentos dos indivíduos e populações. Estas informações determinam a dinâmica de ocupação e dispersão destas espécies (SOBERÓN; NAKAMURA, 2009).

Nestes estudos de caso serão aplicados os modelos de nicho ecológico grinelliano, que relacionam variáveis ambientais com a probabilidade de ocorrência de espécies, para predizer a presença, ausência ou abundância da espécie em uma região (HIRZEL; LE LAY, 2008; SOBERÓN, 2007).

5.1 Processo de Modelagem de Distribuição de Espécies

O processo da MDE pode ser composto por variadas etapas, dependendo dos objetivos da análise e dos dados disponíveis, mas que geralmente contemplam a pré-análise, análise e pós-análise, na qual cada uma destas etapas possuem diversas atividades (CORRÊA et al., 2011; ELITH; LEATHWICK, 2009).

O processo de MDE descrito por Santana et al. (2008) foi adaptado na forma de um diagrama e é apresentado na Figura 27. As etapas que compõem este processo são apresentadas a seguir:

Figura 27 – Etapas que compõem o processo de modelagem de distribuição de espécies.

Fonte: Autor.

- **Definição do problema:** Consiste na definição do experimento de modelagem, que inclui a especificação do conjunto de perguntas que deverão ser respondidas pelo modelo, com base nas espécies selecionadas, a área a ser estudada, e a resolução geográfica das camadas ambientais;
- **Tratamento dos pontos de ocorrência:** Os pontos (registros) de ocorrência e ausência (quando existentes) são obtidos, por meio da coleta em campo ou pela consolidação de registros provenientes de bases de dados e publicações científicas. Em seguida, estes dados são preparados para o experimento de modelagem. Algumas atividades comuns nesta etapa são: padronização, correção, verificação de *outliers*, clusterização e limpeza de registros duplicados;
- **Tratamento de dados ambientais:** Os dados ambientais são obtidos e preparados para a modelagem, passando por atividades de padronização, ajuste de formatos, sistemas de coordenadas, projeções e resoluções. Geralmente estes dados são formatados em arquivos *raster*;
- **Análise da viabilidade dos dados:** São avaliados os dados obtidos nas três etapas anteriores, juntamente com as perguntas do experimento, e

decidido se estes dados são suficientes e adequados para a continuidade do processo de modelagem;

- **Escolha do algoritmo:** São escolhidos os algoritmos que serão utilizados para a geração do modelo de nicho ecológico, com base nos dados disponíveis. Diferentes algoritmos possuem diferentes parâmetros e requisitos (TOWNSEND PETERSON; PAPEŞ; EATON, 2007);
- **Definição dos parâmetros:** São determinados os parâmetros de entrada que irão definir o comportamento do algoritmo de modelagem durante sua execução. O funcionamento do algoritmo depende destes parâmetros, que influenciam diretamente no resultado;
- **Geração do modelo:** O experimento é executado utilizando os dados que foram coletados e preparados e os parâmetros definidos para os algoritmos;
- **Pós-análise:** Após a geração o modelo é avaliado, com o objetivo de verificar a sua adequabilidade à espécie alvo. Técnicas estatísticas podem ser usadas nesta avaliação;
- **Validação pelos pesquisadores:** Com base na avaliação de um pesquisador especialista na espécie, juntamente com os resultados estatísticos da validação do modelo, é decidido se o modelo é válido ou se o processo deve ser refeito.

Conforme apresentado, o processo de modelagem de distribuição de espécies é complexo, pois possui diversas atividades e exige a utilização de um grande volume de dados, além de possibilitar a utilização de uma ampla variedade de ferramentas e algoritmos (PENNINGTON et al., 2007; SANTANA et al., 2008).

O registro da proveniência dos modelos gerados, juntamente com o registro das decisões realizadas durante o processo de modelagem, é de fundamental importância para a validação e reuso dos resultados do experimento, bem como para a compreensão das estratégias empregadas, das limitações nos dados e eventuais vieses do experimento.

A partir da utilização do protótipo funcional apresentado na seção 4.2.5, foi possível a coleta, armazenamento e compartilhamento na Web dos metadados de proveniência gerados nos dois experimentos realizados.

5.2 Estudo de Caso 1

5.2.1 Contextualização

O objetivo deste estudo de caso foi validar a arquitetura computacional para gestão de metadados de proveniência, a partir da coleta, armazenamento e disponibilização dos metadados gerados durante um experimento de modelagem de nicho ecológico.

O experimento, que teve sua proveniência registrada, buscou analisar a relação entre os aspectos físicos e as ocorrências de uma espécie endêmica da região semiárida brasileira, o juazeiro (*Ziziphus juazeiro*). A partir da geração deste modelo de distribuição, foi possível identificar as áreas com características ideais para a adequabilidade da espécie.

Neste estudo de caso foi considerado uma versão preliminar do perfil de aplicação BioProv, descrito no Capítulo 3. A Figura 28 demonstra o modelo de dados considerado no experimento.

Figura 28 – Diagrama do modelo de dados de proveniência considerado no estudo de caso 1.

Fonte: Autor.

Este experimento e seus resultados foram publicados em (SILVA et al., 2016).

5.2.2 Descrição do experimento

O experimento foi realizado a partir da criação de um script na linguagem R (RDC TEAM, 2006), disponível no Apêndice C.

O script realiza a coleta dos registros de ocorrência do *Ziziphus juazeiro* na base de dados do GBIF, por meio do pacote *rgbif* (CHAMBERLAIN et al., 2014). Os dados ambientais utilizados são provenientes da base de dados do Worldclim 1.4 (HIJMANS et al., 2005). Neste experimento foram consideradas as seguintes variáveis ambientais:

- BIO1: Temperatura média anual;
- BIO5: Temperatura máxima do mês mais quente;
- BIO6: Temperatura mínima do mês mais frio;
- BIO7: Variação anual de temperatura;
- BIO8: Temperatura média do trimestre mais chuvoso;
- BIO12: Precipitação anual;

- BIO16: Precipitação do trimestre mais chuvoso;
- BIO17: Precipitação do trimestre mais seco.

Para a modelagem foi utilizado a implementação do algoritmo Bioclim (BOOTH et al., 2014; HIJMANS; GRAHAM, 2006), disponível no pacote *dismo* (HIJMANS et al., 2016).

Para o registro da proveniência foi utilizado o pacote *BioProv Client*, que disponibiliza um conjunto de funções para a descrição, armazenamento e consulta dos metadados de proveniência. A Figura 29 apresenta o processo considerado na execução deste estudo de caso.

Figura 29 – Etapas do estudo de caso 1.

Fonte: Autor.

Na etapa 1, durante a execução do experimento de modelagem, os metadados descritos no *script* foram enviados para o servidor de proveniência, por meio dos seus serviços Web. Este processo é transparente para o pesquisador, que precisa apenas descrever os metadados utilizando as funções disponibilizadas pelo pacote *BioProv Client*.

A Figura 30 apresenta a função *createAgentProvenance()* utilizada para descrever os dois pesquisadores envolvidos no experimento. O parâmetro *uri*

representa o identificador único do pesquisador, enquanto que os demais parâmetros são as propriedades e relações do pesquisador, conforme as regras definidas no perfil de aplicação BioProv.

Figura 30 – Descrição dos metadados dos pesquisadores envolvidos no experimento.

```
ag_daniel <- createAgentProvenance(uri = 'http://bioprov.poli.usp.br/agents/6613529',
                                   type = 'Researcher',
                                   name = 'Daniel Lins da Silva',
                                   identifier = 'http://bioprov.poli.usp.br/agents/6613529',
                                   mbox = 'mailto:daniellins@usp.br',
                                   seeAlso = 'http://lattes.cnpq.br/6502450550040226')

ag_andre <- createAgentProvenance(uri = 'http://bioprov.poli.usp.br/agents/8788438',
                                   type = 'Researcher',
                                   name = 'Andre Filipe de Moraes Batista',
                                   identifier = 'http://bioprov.poli.usp.br/agents/8788438',
                                   mbox = 'mailto:andrefmb@usp.br',
                                   seeAlso = 'http://lattes.cnpq.br/2015345575533420')
```

Fonte: Autor.

Na etapa 2, ao receber as requisições do *BioProv Client*, o servidor de proveniência valida, processa e converte os metadados. Em seguida, na etapa 3, estes metadados são armazenados no banco de grafos RDF.

A Figura 31 apresenta a função *createActivityProvenance()* utilizada para descrever a atividade de modelagem deste experimento.

Figura 31 – Descrição da atividade de modelagem do experimento.

```
act_dataModeling <- createActivityProvenance(bundle = bundle,
                                             uri = 'http://bioprov.poli.usp.br/analysing',
                                             type = 'Analysis',
                                             title = 'Species Distribution Modeling with the Bioclim Algorithm',
                                             used = list(e_dataset, e_environmentalLayers, e_enmAlgorithm),
                                             generated = e_resultModeling)
```

Fonte: Autor.

Todas as funções de descrição de metadados possuem o parâmetro opcional *bundle*. O *bundle* corresponde a um agrupamento de metadados. Ao criar um *bundle* e associar sua URI nas descrições de metadados, automaticamente estas descrições estão sendo agrupadas. O *bundle* no contexto deste estudo de caso representa o experimento científico.

O parâmetro *used* identifica os recursos consumidos pela atividade. Nesta atividade foram utilizados um *dataset* de ocorrências da espécie *Ziziphus juazeiro*, as camadas ambientais provenientes do Wordclim e o algoritmo de modelagem Bioclim.

O parâmetro *generated*, identifica os recursos gerados pela atividade. Nesta atividade o modelo de distribuição da espécie foi gerado.

Durante o processamento destes metadados, o parâmetro *used* gera relações *prov:used* para cada um dos recursos descritos, e o parâmetro *generated* gera uma relação *prov:wasGeneratedBy*. A Figura 32 apresenta o grafo dos metadados descritos na atividade de modelagem do experimento.

Figura 32 – Representação do grafo de proveniência da atividade de modelagem do experimento.

Fonte: Autor.

Na etapa 4, após a finalização do experimento, foi gerado o relatório de proveniência do mapa de distribuição da *Ziziphus Juazeiro*, que demonstra todos os metadados associados a este recurso de dados. A Figura 33 apresenta o trecho do *script* no qual foram requisitados dois relatórios de proveniência do mapa gerado pelo experimento, um gráfico e um textual.

O pacote *BioProv Client* disponibiliza duas funções para geração de relatórios de proveniência. A *getProvenanceOfEntity()* e a *getProvenanceOfBundle()*. A primeira função, recebe como parâmetro a URI de um recurso de dados e gera um relatório de proveniência com todos os metadados associados ao recurso. A segunda função,

recebe como parâmetro a URI de um experimento (*bundle*) e gera o relatório de proveniência com todos os metadados agrupados dentro do experimento.

Figura 33 – Função do pacote BioProv Client utilizada na requisição de dois relatórios sobre a proveniência do experimento.

```
getProvenanceOfEntity(uri = e_FinalresultModeling, format = 'png')
getProvenanceOfEntity(uri = e_FinalresultModeling, format = 'rdf')
```

Fonte: Autor.

As duas funções também possuem um parâmetro opcional *format*, que define o formato do relatório, que pode ser JSON-LD, RDF (Turtle), PNG ou SVG. Os relatórios de proveniência gerados no experimento estão disponíveis nos Apêndices D e E.

Na etapa 5, foi gerado o relatório de proveniência do mapa de distribuição da *Ziziphus Juazeiro* diretamente no navegador Web, a partir do *endpoint* <http://bioprov.poli.usp.br/api/provenance/>. Como parâmetro para o serviço foi informado a URI <http://bioprov.poli.usp.br/models/ziziphusJuazeiro>, que identifica o recurso desejado. A Figura 34 apresenta o relatório de proveniência gerado após a requisição do serviço.

Figura 34 – Relatório de proveniência do mapa de distribuição da *Ziziphus Juazeiro* gerado no navegador Web.

```
{
  "@context": {
    "arctos.database.museum": "http://arctos.database.museum/",
    "bioprov": "http://bioprov.poli.usp.br/",
    "cefa.dri.edu": "http://www.cefa.dri.edu/",
    "cran.r-project.org": "http://cran.r-project.org/",
    "dbpedia": "http://dbpedia.org/",
    "dcterms": "http://purl.org/dc/terms/",
    "desp.ucdavis.edu": "http://desp.ucdavis.edu/",
    "doi": "http://dx.doi.org/",
    "dwc": "http://rs.tdwg.org/dwc/terms/",
    "foaf": "http://xmlns.com/foaf/0.1/",
    "github": "https://github.com/",
    "gitlab": "https://gitlab.com/",
    "homepages.inf.ed.ac.uk": "http://homepages.inf.ed.ac.uk/",
    "ib.usp": "http://ib.usp.br/",
    "manisnet.org": "http://manisnet.org/",
    "orcid": "http://orcid.org/",
    "poli": "http://poli.usp.br/",
    "prism.oregonstate.edu": "http://www.prism.oregonstate.edu/",
    "prov": "http://www.w3.org/ns/prov#",
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "ropensci.org": "http://ropensci.org/",
    "rstudio": "http://www.rstudio.com/",
    "sensor.com": "http://sensor.com/",
    "teste": "http://teste.com/",
    "time": "http://www.w3.org/2006/time#",
    "tools.com": "http://tools.com/",
    "uc-ciee": "http://www.uc-ciee.org/",
    "ucdavis.edu": "http://desp.ucdavis.edu/",
    "usp": "http://usp.br/",
    "worldclim": "http://www.worldclim.org/",
    "xsd": "http://www.w3.org/2001/XMLSchema"
  },
  "@graph": [
    {
      "@id": "http://bioprov.poli.usp.br/entities/607bf17b-b55a-4c73-a212-c03e964d77b9",
      "@type": "prov:Entity"
    }
  ]
}
```

Fonte: Autor.

O relatório de proveniência foi gerado no formato JSON-LD. No entanto, utilizando o parâmetro *format* é possível a alteração do formato do relatório para RDF (Turtle), PNG ou SVG. A Figura 35 apresenta o relatório no formato PNG.

Figura 35 - Relatório de proveniência do mapa de distribuição da *Ziziphus Juazeiro* gerado no formato PNG.

Fonte: Autor.

Em seguida, o mapa de distribuição gerado foi publicado em uma página Web para a visualização da sua proveniência por meio do *plug-in* BioProv.

Com o *plug-in* BioProv ativado no navegador, todas as páginas Web abertas são analisadas. Caso seja identificado na página HTML um recurso de dados com proveniência disponível, automaticamente o *plug-in* consulta os metadados deste recurso. O resultado da consulta é disponibilizado em uma nova aba do navegador.

A Figura 36 apresenta a página Web no qual o recurso foi publicado e os resultados do relatório de proveniência gerados pelo *plug-in* BioProv.

Figura 36 – À esquerda, a página Web no qual o mapa do modelo de distribuição foi publicado. À direita, o relatório de proveniência gerado pelo *plug-in* BioProv.

Fonte: Autor.

O *plug-in* BioProv analisa os metadados publicados na página Web e identifica os recursos de dados que possuem proveniência, com base nas recomendações do PROV (MOREAU et al., 2013). A Figura 37 apresenta os metadados incluídos na página Web considerada no experimento.

Figura 37 – Metadados incluídos na página Web para identificar a existência de um recurso com proveniência disponível para consulta.

```
<link rel="http://www.w3.org/ns/prov#has_query_service" href="http://bioprov.poli.usp.br/api/provenance/" />
<link rel="http://www.w3.org/ns/prov#has_anchor" href="http://bioprov.poli.usp.br/models/ziziphusJoazeiro" />
```

Fonte: Autor.

5.2.3 Discussões

Os resultados demonstraram que a arquitetura computacional proposta é adequada para a gestão dos metadados de proveniência de um experimento científico. Os processos de descrição, armazenamento e compartilhamento destes metadados foram concluídos sem a ocorrência de problemas ou limitações.

Por meio do pacote *BioProv Client* foi possível o registro dos metadados de proveniência em um nível de granularidade adequado para o experimento. No entanto, foi identificado que a versão utilizada do perfil de aplicação BioProv necessitava de elementos mais específicas para a descrição eficiente dos recursos de dados do

experimento. Por isso, um aspecto a ser considerado é o melhoramento do perfil de aplicação BioProv.

A estratégia do pesquisador descrever os metadados de proveniência diretamente no *script* do experimento demonstrou vantagens, como a flexibilidade na definição do nível de detalhes dos metadados. Possibilitando que apenas as informações essenciais para a reprodutibilidade do experimento sejam descritas. Outra vantagem é que no reuso dos *scripts* para novos experimentos as descrições dos metadados de proveniência também poderão ser reutilizadas, diminuindo o esforço na sua preparação.

No entanto, o processo de descrição dos metadados de proveniência exige um esforço significativo. Principalmente em processos complexos e que exigem metadados detalhados sobre o processo. Por isso, um segundo aspecto a ser considerado é a análise de estratégias que exijam um menor esforço para a descrição dos metadados.

Os metadados enviados pelo *BioProv Client*, durante a execução do experimento, foram corretamente convertidos e processados pelo servidor de proveniência, com base no perfil de aplicação BioProv. Após a conversão destes dados em grafos RDF, os mesmos foram registrados no repositório de metadados.

Os metadados de proveniência do experimento tornaram-se disponíveis na Web imediatamente após o seu armazenamento. O *BioProv Client* e o *plug-in* BioProv geraram relatórios de proveniência de recursos de dados e do experimento completo por meio dos seus identificadores únicos.

A visualização dos relatórios de proveniência foi disponibilizada em quatro formatos: JSON-LD, RDF, PNG e SVG. Esta variedade de formatos possibilita que a arquitetura computacional seja considerada por diferentes sistemas. O JSON-LD possibilita que os sistemas compatíveis com JSON possam conectar-se aos serviços de proveniência. Já os sistemas compatíveis com a tecnologias da Web Semântica, podem consumir os dados no formato RDF.

Por fim, a visualização da proveniência nos formatos gráficos é voltada para os pesquisadores, que poderão analisar os metadados de proveniência rapidamente. No entanto, experimentos complexos geram grafos muito grandes e proporcionalmente complexos, dificultando sua análise e compreensão. Por isso, o terceiro aspecto a ser considerado é a análise de novas técnicas para a visualização eficiente de grandes grafos de proveniência.

5.3 Estudo de Caso 2

5.3.1 Contextualização

O segundo estudo de caso apresentado considerou dois objetivos principais. O primeiro objetivo foi validar o perfil de aplicação BioProv, descrita no Capítulo 3. Com base nos experimentos anteriores, este PA passou por atualizações, tanto na sua estrutura de classes e relações, quanto nas propriedades disponíveis. O segundo objetivo foi validar o serviço de consultas da arquitetura computacional, que permite que pesquisadores e outros sistemas possam fazer consultas personalizadas no repositório de metadados de proveniência, otimizando o processo de extração e análise destes metadados.

O experimento, que teve sua proveniência registrada, foi desenvolvido dentro do escopo do projeto de pesquisa “*Testing methods for predicting mammalian specie’s records of response to 20th century change in California*” (SMITH; ROWE, 2011). O objetivo principal deste projeto foi compreender como as recentes mudanças climáticas estão afetando os mamíferos na Califórnia. Este projeto gerou artigos e relatórios que descrevem, com um bom nível de detalhes, os experimentos realizados. Os dados disponibilizados nestas publicações possibilitaram a reprodutibilidade do experimento e a descrição dos seus metadados de proveniência. Este estudo de caso considerou a MDE de um esquilo da espécie *Uroditellus beldingi*.

5.3.2 Descrição do Experimento

O experimento seguiu a mesma estratégia do primeiro estudo de caso, e foi realizado a partir da criação de um *script* na linguagem R, disponível no Apêndice F.

O script realiza a coleta dos registros de ocorrência do *Uroditellus beldingi* de duas bases de dados, o Portal Manis (<http://manisnet.org/>) (STEIN; WIECZOREK, 2004) e o Portal Arctos (<http://arctos.database.museum>).

Os dados ambientais, foram obtidos do portal do *PRISM Climate Group* (<http://www.prism.oregonstate.edu/>). Neste experimento foram consideradas as seguintes variáveis ambientais:

- BIO2: Média da variação diurna de temperatura;
- BIO3: Isotermalidade;
- BIO5: Temperatura máxima do mês mais quente;
- BIO6: Temperatura mínima do mês mais frio;
- BIO7: Variação anual de temperatura;
- BIO13: Precipitação do mês mais chuvoso;
- BIO14: Precipitação do mês mais seco;
- BIO15: Sazonalidade de precipitação;
- BIO18: Precipitação do trimestre mais frio.

O algoritmo utilizado para a modelagem foi o *Maximum Entropy* (MaxEnt) (PHILLIPS et al., 2017; PHILLIPS; ANDERSON; SCHAPIRE, 2006). Para executá-lo no R foi utilizado o pacote *dismo*, que disponibiliza uma interface para a sua configuração e execução.

Para o registro da proveniência foi utilizado o pacote *BioProv Client*, que também passou por atualizações, juntamente com os serviços de proveniência. A descrição da versão atualizada dos serviços de proveniência está detalhada no Capítulo 4.

Após a descrição de todo o experimento e a sua execução, foram realizadas consultas SPARQL para a extração e análise dos metadados de proveniência gerados

durante o experimento. A Figura 38 apresenta o processo considerado na execução deste estudo de caso.

Figura 38 – Etapas do processo do estudo de caso 2.

Fonte: Autor.

As etapas de 1 a 3 foram semelhantes as etapas realizadas no primeiro estudo de caso.

Na etapa 4, foram gerados os relatórios de proveniência. Primeiramente foi gerado o relatório de proveniência do experimento (<http://bioprov.poli.usp.br/experiment/31>), por meio da função *getProvenanceOfExperiment()*. Este relatório é apresentado na Figura 39.

Figura 39 – Relatório de proveniência do experimento realizado no estudo de caso 2.

Fonte: Autor.

O relatório de proveniência do experimento disponibilizou informações sobre o seu contexto. Este experimento foi realizado por dois pesquisadores, Daniel Lins e André Batista, foi derivado de um relatório, de 2011, do projeto “*Testing methods for predicting mammalian specie’s records of response to 20th century change in California*”, que foi coordenado pelo pesquisador Allan B. Smith.

Em seguida foram gerados os relatórios de proveniência do processo do experimento. Os relatórios em PNG e RDF estão disponíveis nos Apêndices G e H, respectivamente.

Na etapa 5, foi validado o SPARQL *endpoint*, disponibilizado juntamente com os serviços de proveniência, para que pesquisadores e outros sistemas pudessem realizar consultas e análises personalizadas no repositório de metadados de proveniência.

O SPARQL *endpoint* foi disponibilizado a partir do endereço <http://bioprov.poli.usp.br/api/sparql>. Para o envio das consultas foi necessário a codificação da consulta SPARQL e seu envio pelo parâmetro *query*. Estas características estão de acordo com a recomendação da W3C sobre o protocolo

SPARQL 1.1 (FEIGENBAUM et al., 2013). A Figura 40 demonstra uma consulta SPARQL e como essa consulta foi enviada ao SPARQL *endpoint*.

Figura 40 – Consulta SPARQL e sua requisição ao SPARQL *endpoint* do Servidor de Proveniência.

Consulta SPARQL

```
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?agent ?name
WHERE {
  ?agent a prov:Agent.
  ?agent foaf:name ?name.
}
```

Requisição no SPARQL endpoint do Servidor de Proveniência BIOPROV

```
http://bioprov.poli.usp.br/api/sparql?query=PREFIX%20prov%3A%20%20%3Chttp%3A%2F%2Fwww.w3.org%2Fns%2Fprov%23%3E%20PREFIX%20foaf%3A%20%20%3Chttp%3A%2F%2Fxmlns.com%2Ffoaf%2F0.1%2F%3E%20%20SELECT%20%3Fagent%20%3Fname%20WHERE%20%7B%20%20%20%20%20%3Fagent%20a%20prov%3AAgent.%20%20%20%20%20%3Fagent%20foaf%3Aname%20%3Fname.%20%7D
```

Fonte: Autor.

Para os testes do SPARQL *endpoint* e a validação do perfil de aplicação BioProv, foram definidas questões sobre o experimento científico para serem respondidas a partir dos metadados de proveniência armazenados. As questões e suas respostas serão descritas a seguir.

Questão 1. Quais as atividades executadas durante o experimento?

Figura 41 – Consulta SPARQL da questão 1, do estudo de caso 2.

```
PREFIX prov: <http://www.w3.org/ns/prov#>

SELECT ?activity ?when
FROM <http://bioprov.poli.usp.br/experiment/31>
WHERE {
  ?activity a prov:Activity.
  ?activity dct:created ?when.
}
ORDER BY ?when
```

Fonte: Autor.

Tabela 7 – Resultado da consulta SPARQL da questão 1, do estudo de caso 2.

Atividades	Início
<http://bioprov.poli.usp.br/step1_retrieving>	10/02/2017 03:02:28
<http://bioprov.poli.usp.br/step2_retrieving>	10/02/2017 03:02:28
<http://bioprov.poli.usp.br/step3_merging>	10/02/2017 03:02:28
<http://bioprov.poli.usp.br/step4_cleaning>	10/02/2017 03:02:28
<http://bioprov.poli.usp.br/step5_analysing>	10/02/2017 03:04:28
<http://bioprov.poli.usp.br/step6_prediction>	10/02/2017 03:04:36

Fonte: Autor.

Questão 2. Qual a origem dos registros de ocorrência da espécie *Urocyon beldingi*? Como foram obtidos? Quantos registros foram utilizados?

Figura 42 – Consulta SPARQL da questão 2, do estudo de caso 2.

```

PREFIX prov:    <http://www.w3.org/ns/prov#>
PREFIX bioprov: <http://bioprov.poli.usp.br/>
SELECT ?activity ?datasetSource ?datasetQuantity ?inputType ?inputName
       ?inputValue
FROM <http://bioprov.poli.usp.br/experiment/31>
WHERE {
    ?activity a prov:Activity.
    OPTIONAL{ ?activity bioprov:activityType "Acquisition".
               ?activity prov:used ?input.
               ?input dcterms:identifier ?inputName.
               ?input bioprov:entityType ?inputType.
               OPTIONAL {?input rdf:value ?inputValue.}
               OPTIONAL {?input dcterms:source ?inputValue.}
            }
    OPTIONAL{ ?activity bioprov:activityType "Preparation".}
    ?dataset prov:wasGeneratedBy ?activity.
    ?dataset bioprov:numberOfRecords ?datasetQuantity.
    ?dataset dcterms:source ?datasetSource.
}
ORDER BY ?activity ?inputType

```

Fonte: Autor.

Tabela 8 – Resultado da consulta SPARQL da questão 2, do estudo de caso 2.

Atividades	Tipo	Nome	Valor
<http://bioprov.poli.usp.br/step1_retrieving>	InputParameter	endDate	2009
		startDate	1970
		species	Urocitellus beldingi
	Software	ARCTOS	http://arctos.database.museum
	Dataset gerado		Quantidade de registros
	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/speciesRecords/Urocitellus_beldingi_Arctos.csv		111
<http://bioprov.poli.usp.br/step2_retrieving>	InputParameter	endDate	2009
		startDate	1970
		species	Urocitellus beldingi
	Software	MANIS	http://manisnet.org/
	Dataset gerado		Quantidade de registros
	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/speciesRecords/Urocitellus_beldingi_Manis.txt		174
<http://bioprov.poli.usp.br/step4_cleaning>	Dataset gerado		
	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/speciesRecords/Urocitellus_beldingi_training_modern_xy.csv		32

Fonte: Autor.

Questão 3. Quais os critérios considerados na limpeza e preparação dos registros de ocorrência utilizados na modelagem?

Figura 43 – Consulta SPARQL da questão 3, do estudo de caso 2.

```
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX bioprov: <http://bioprov.poli.usp.br/>

SELECT ?activity ?title ?source
FROM <http://bioprov.poli.usp.br/experiment/31>
WHERE {
    ?activity a prov:Activity.
    ?activity bioprov:activityType "Preparation".
    ?activity dcterms:title ?title.
    ?activity dcterms:description ?source.
}
```

Fonte: Autor.

Tabela 9 – Resultado da consulta SPARQL da questão 3, do estudo de caso 2.

Cr�terios para limpeza e prepara��o dos dados
1. Retained records from the entire conterminous US west of the eastern border of the Rocky Mountains
2. Retained records With coordinate uncertainty >5000 m
3. Removed records that were below the 0.25th and above the 99.75th percentiles in either mean annual temperature or precipitation relative to the other sites where the species was located
4. Removed records collected before year 2000 with coordinate uncertainties <3 m
5. Removed records from 2000 onward with coordinate uncertainties >200 m
6. Removed records outside the known species distribution area;
7. Records were thinned so that no presence points were within 1km of one another

Fonte: Autor.

Questo 4. Qual o algoritmo utilizado na modelagem? Quais os parmetros de entrada e recursos utilizados? Qual o tempo de execuo do algoritmo?

Figura 44 – Consulta SPARQL da questo 4, do estudo de caso 2.

```

PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX bioprov: <http://bioprov.poli.usp.br/>
PREFIX time: <http://www.w3.org/2006/time#>

SELECT ?activity ?duration ?comput ?inputName ?inputType ?inputSource
?inputVersion ?inputValue
FROM <http://bioprov.poli.usp.br/experiment/31>
WHERE {
    ?activity a prov:Activity.
    ?activity bioprov:activityType "Analysis".
    ?activity time:hasDuration ?duration.
    ?activity bioprov:computationalArchitectureUsed ?comput.
    ?activity prov:used ?input.
    optional {?input bioprov:entityType ?inputType.}
    optional {?input dcterms:title ?inputName.}
    optional {?input dcterms:identifier ?inputName.}
    optional {?input dcterms:source ?inputSource.}
    optional {?input dcterms:isVersionOf ?inputVersion.}
    optional {?input rdf:value ?inputValue.}
}

```

Fonte: Autor.

Tabela 10 – Resultado da consulta SPARQL da questão 4, do estudo de caso 2.

Atividade	Duração	Equipamento utilizado notebook core i5, 8Gb RAM, SSD 256Gb		
step5_analysing	1.877686 min			
Tipo de recurso utilizado	Nome	Origem	Valor	Versão
Software	RStudio Software	http://www.rstudio.com/products/rstudio/download/		0.99.893
software	dismo: Species Distribution Modeling	https://cran.r-project.org/src/contrib/dismo_1.1-1.tar.gz		
Software	MAXENT Software	https://www.cs.princeton.edu/~schapire/maxent/		3.3.3
InputParameter	askoverwrite		false	
InputParameter	jackknife		true	
InputParameter	responsecurves		true	
InputParameter	threads		2	
InputParameter	hinge		true	
InputParameter	threshold		true	
InputParameter	product		true	
InputParameter	quadratic		true	
InputParameter	linear		true	
InputParameter	betamultiplier		1	
InputParameter	randomtestpoints		30	
DataResource	Mean diurnal temperature range	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO02.asc		
DataResource	Isothermality (BIO2/BIO07)	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO03.asc		
DataResource	Maximum temperature of the warmest month	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO05.asc		
DataResource	Minimum temperature of the coldest month	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO06.asc		
DataResource	Temperature annual range (BIO05-BIO06)	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO07.asc		
DataResource	Precipitation of Wettest Month	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO13.asc		
DataResource	Precipitation of Driest Month	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO14.asc		
DataResource	Precipitation Seasonality (Coefficient of Variation)	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO15.asc		
DataResource	Precipitation of Warmest Quarter	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO18.asc		
DataResource	Occurrence file of the <i>Urocitellus_beldingi</i> species	https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/speciesRecords/Urocitellus_beldingi_training_modern_xy.csv		

Fonte: Autor.

Questão 5. Qual a estratégia para a escolha das variáveis ambientais utilizadas na modelagem?

Figura 45 – Consulta SPARQL da questão 5, do estudo de caso 2.

```
PREFIX prov:      <http://www.w3.org/ns/prov#>
PREFIX bioprov:  <http://bioprov.poli.usp.br/>
SELECT ?activity ?decisions
FROM <http://bioprov.poli.usp.br/experiment/31>
WHERE {
    ?activity a prov:Activity.
    ?activity bioprov:activityType "Analysis".
    ?activity bioprov:decisions ?decisions.
}
```

Fonte: Autor.

Tabela 11 – Resultado da consulta SPARQL da questão 5, do estudo de caso 2.

Atividade	Decisões
<http://bioprov.poli.usp.br/step5_analysing>	From 19 BIOCLIM variables, were kept those are expected to be biologically meaningful to the species and that had cross-correlations >-0.7 and <0.7 (Tingley and Herman 2009)

Fonte: Autor.

5.3.3 Discussões

Os resultados demonstraram que a atualização do perfil de aplicação BioProv permitiu uma descrição mais detalhada e padronizada da proveniência do experimento científico. Este resultado foi comprovado também com os resultados obtidos a partir das consultas realizadas no SPARQL *endpoint*.

A disponibilização de um SPARQL *endpoint*, para a realização de consultas, mostrou-se uma ferramenta poderosa para a análise e a descoberta de informações na base de metadados de proveniência. Além disso, com o uso do perfil de aplicação BioProv, consultas geradas para um experimento poderão ser reutilizadas em novos experimentos, devido a padronização do grafo de proveniência.

5.4 Considerações finais

Os dois estudos de casos, gerados em momentos diferentes da pesquisa, apresentaram resultados positivos.

No primeiro estudo de caso, a arquitetura de coleta, armazenamento e publicação dos dados de proveniência mostrou-se adequada e eficiente no processo de gestão dos metadados de proveniência de um experimento científico. Por meio de um identificador único, definido pelo pesquisador ou gerado pela arquitetura computacional, uma pessoa ou sistema pode ter acesso a proveniência de um experimento ou conjunto de dados que teve sua proveniência registrada. Além disso, não foram verificados impactos significativos de desempenho, durante a execução do experimento, devido os mecanismos de registro da proveniência.

No segundo estudo de caso, a atualização do perfil de aplicação permitiu uma descrição mais precisa do experimento, contemplando as informações vitais para a compreensão, reprodutibilidade e reutilização dos produtos gerados. Além disso, por meio das consultas SPARQL, foi possível a extração e análise de diversas informações sobre o experimento. Pôde-se comprovar que apesar do repositório de proveniência armazenar uma grande quantidade de informações, que torna a análise do grafo de proveniência bastante trabalhosa, o uso das consultas SPARQL reduziu significativamente este esforço.

Novas aplicações baseadas em SPARQL, com foco na análise dos metadados de proveniência, poderão tornar o processo de análise e visualização da proveniência de recursos de dados adequadas tanto para pesquisadores quanto para gestores.

6 CONCLUSÃO

O objetivo geral da pesquisa foi apresentar a aplicabilidade de uma estratégia computacional para apoiar a reprodutibilidade e o reuso dos dados científicos por meio da gestão de metadados de proveniência. Esta estratégia considerou uma solução

aplicada a ambientes distribuídos e heterogêneos, devido as características da área da Informática para Biodiversidade.

Para o desafio da reprodutibilidade e reuso dos dados científicos, ocasionado principalmente pela ausência de informações que apoiem a compreensão e validação dos dados e experimentos, foi proposto no Capítulo 4 uma arquitetura computacional para gerenciar metadados de proveniência durante o ciclo de vida dos dados. Estes metadados, mesmo descritos em momentos distintos e por diferentes ferramentas, são conectadas em um grafo de proveniência que relaciona as atividades, pessoas e recursos envolvidos na geração e evolução de um recurso de dados.

Posteriormente, todos estes metadados, que descrevem a origem de um recurso podem ser recuperados na Web por meio do seu identificador único.

A utilização de identificadores únicos em dados é um tema de grande interesse na comunidade científica. Diversas pesquisas e iniciativas buscam definir estratégias para a popularização do uso de identificadores únicos nas bases de dados científicos (GURALNICK et al., 2015; PAGE, 2008; VAN DE SOMPEL et al., 2014). A arquitetura computacional proposta possibilita que os identificadores únicos dos recursos de dados possam ser utilizados também para o registro da sua proveniência, como um DOI associado a um conjunto de dados, possibilitando que estas bases de dados sejam conectadas.

Para o registro da proveniência de um recurso de dados que não possui um identificador único associado, a arquitetura computacional realiza a geração de identificadores únicos seguindo o padrão HTTP URI. Assim, da mesma forma que um DOI possibilita a recuperação dos metadados e do arquivo de uma publicação científica, o identificador único de proveniência possibilita a recuperação dos metadados que descrevem a história daquele recurso de dados.

Conforme apresentado nos Capítulos 1 e 2, as estratégias mais populares no uso da proveniência dos dados são as associadas aos WfMS. No entanto, estas soluções geram uma dependência de uso destas ferramentas, além de não considerarem seu uso em ambientes distribuídos. O desafio de criar uma estratégia

de proveniência adequada para ambientes heterogêneos e distribuídos foi solucionado por meio da especificação de uma arquitetura computacional com componentes desacoplados para a descrição e aquisição, gestão e armazenamento dos metadados de proveniência. Com isso, o componente de aquisição dos metadados de proveniência pode ser um pacote R, uma ferramenta Web e até mesmo um WfMS. Independentemente da ferramenta utilizada para a aquisição dos metadados de proveniência, o armazenamento destes metadados é realizado por meio dos serviços web disponibilizados pelo servidor de proveniência.

Os requisitos computacionais para o uso dos serviços de proveniência são apenas dois: compatibilidade com requisições HTTP e o uso de JSON na serialização dos metadados. Estes dois requisitos são compatíveis com os sistemas computacionais atuais, e utilizam os principais padrões da Web. Com isso, sistemas gerenciadores de coleções biológicas, sistemas de qualidade e curadoria, ferramentas de análise de dados e *scripts* em diversas linguagens de programação podem armazenar e consultar metadados de proveniência usando essa arquitetura computacional.

Ainda no contexto dos ambientes heterogêneos e distribuídos, existe a necessidade do processamento e da consolidação dos metadados de proveniência gerados por diferentes ferramentas. Este desafio foi solucionado com a definição de um perfil de aplicação que padroniza como os metadados de proveniência são organizados e armazenados. Os detalhes do perfil de aplicação BioProv são descritos no Capítulo 3. Além dos benefícios do registro e armazenamento padronizado dos metadados, a utilização de um perfil de aplicação possibilita que outras ferramentas possam extrair e analisar estes metadados, visto que os grafos de proveniência disponíveis seguem um padrão na descrição das suas relações e propriedades.

Por fim, outro desafio considerado neste trabalho foi a disponibilização na Web dos metadados de proveniência, para que pesquisadores pudessem consultar a proveniência de recursos de dados e experimentos científicos, apoiando a compreensão, reprodutibilidade e o reuso destes dados. Para isso, a arquitetura computacional disponibilizou serviços Web para a geração de relatórios de

proveniência, que podem ser acessados por sistemas ou por pesquisadores diretamente no navegador Web. Estes serviços necessitam apenas do identificador único do recurso de dados ou do experimento científico para serem gerados em diversas formatos, como JSON, RDF, PNG e SVG. Além disso, consultas personalizadas podem ser submetidas ao servidor de proveniência por meio de consultas SPARQL, possibilitando a extração e análise de metadados específicos.

Conforme descrito em Wazlawick (2009), os objetivos específicos de uma tese são subprodutos do objetivo geral, não triviais e verificáveis ao final do trabalho. Desse modo, nos itens a seguir, são reproduzidos os objetivos específicos da pesquisa, sendo evidenciados os respectivos resultados obtidos, verificados ao longo do trabalho:

- **Definir um Perfil de Aplicação (PA), baseado em ontologias e vocabulários existentes, para a estruturação, padronização e armazenamento dos metadados de proveniência.** O Perfil de Aplicação (PA) para a descrição da proveniência de dados de biodiversidade, chamado BioProv, considerou a ontologia do PROV na descrição das relações entre dados, atividades e responsáveis envolvidos nos experimentos científicos. Vocabulários populares da Web Semântica como o RDF, RDFS, FOAF e Dublin Core, tiveram propriedades reutilizadas nas diversas classes do PA. O Darwin Core e o EML, vocabulários populares da área de IB, também tiveram propriedades reutilizadas. Conceitos adicionais que não puderam ser reutilizados de vocabulários existentes foram incorporados ao vocabulário local do BioProv. Também foram definidos os dicionários controlados *bioprov:Type*, *bioprov:ActivityType* e *bioprov:Format*, com o objetivo de padronizar o conteúdo de propriedades consideradas essenciais no PA, e garantir a otimização na recuperação destas informações. Este objetivo específico foi atingido a partir do desenvolvimento e validação do BioProv, descrito no Capítulo 3.
- **Definir uma arquitetura computacional para a captura, gestão e compartilhamento na Web dos metadados de proveniência de**

recursos de dados e experimentos científicos, garantindo assim, que estas informações possam ser descobertas e reutilizadas pelos pesquisadores. A arquitetura computacional foi especificada baseada em três requisitos principais, diretamente relacionados aos objetivos da pesquisa:

- Arquitetura computacional independente de ferramentas, possibilitando sua utilização em ambientes heterogêneos;
- Arquitetura computacional distribuída, possibilitando que os componentes da arquitetura possam estar disponíveis em diferentes computadores;
- Disponibilização dos metadados de proveniência na Web, para possibilitar a recuperação e consulta destes dados, apoiando a compreensão, reprodutibilidade e reuso dos dados.

Requisitos funcionais e não funcionais adicionais foram considerados com base na literatura científica, em discussões com pesquisadores da área de IB e a partir da avaliação dos resultados dos experimentos realizados durante a pesquisa. Este objetivo específico foi atingido a partir da especificação e validação da arquitetura computacional por meio de estudos de caso. A especificação da arquitetura computacional é descrita no Capítulo 4.

- **Validar o perfil de aplicação e a arquitetura computacional para gestão de metadados de proveniência em estudos de caso com cenários na área de biodiversidade.** Dentre os experimentos realizados durante a pesquisa, que possibilitaram a melhoria incremental da estratégia computacional proposta e sua validação, dois deles foram descritos como estudos de caso neste trabalho. O primeiro estudo de caso foi direcionado para a validação da arquitetura computacional, e o segundo estudo de caso foi direcionado para a validação do modelo de dados proposto e sua utilização para o apoio a reprodutibilidade e reuso dos dados. Este objetivo específico foi atingido e os dois estudos de caso são descritos no Capítulo 5.

Nas seções seguintes são descritas as contribuições obtidas com esta pesquisa e os trabalhos em andamento e futuros, relacionados a este trabalho.

6.1 Contribuições da tese

Este trabalho apresentou como contribuição a proposição de uma estratégia computacional baseada em metadados de proveniência para o apoio a reprodutibilidade de experimentos científicos e o reuso dos dados produzidos. Entre as principais contribuições da tese, destacam-se:

- Especificação de uma arquitetura computacional distribuída para a gestão dos metadados de proveniência durante o seu ciclo de vida, e a disponibilização destas informações na Web, por meio de um identificador único associado aos dados e ao experimento. Desta forma, garantindo que as informações vitais sobre os experimentos e seus dados sejam compartilhadas e facilmente consultadas por pesquisadores, apoiando a reprodutibilidade e o reuso na área científica;
- Definição do BioProv, um perfil de aplicação que descreve a estrutura e as diretrizes para a organização e armazenamento dos metadados de proveniência na área de IB, possibilitando que diferentes ferramentas e iniciativas possam compartilhar metadados de proveniência;
- Construção de um protótipo funcional da arquitetura computacional especificada neste trabalho, que realiza a descrição, processamento, armazenamento e compartilhamento dos metadados de proveniência. Estas implementações podem ser utilizadas como referência para a construção de novas ferramentas que poderão armazenar ou consumir metadados de proveniência da arquitetura computacional proposta.

6.2 Trabalhos futuros

Durante a realização desta pesquisa, desafios e oportunidades relacionados a proveniência dos dados foram identificados.

Entre estes desafios, alguns trabalhos estão em andamento. O primeiro trabalho objetiva a geração de um índice de veracidade de dados científicos, que considera os metadados de proveniência para a extração de informações que serão utilizadas no cálculo deste índice. Este trabalho é realizado por André Filipe de Moraes Batista. O segundo trabalho é a modificação da ferramenta openModeller (MUÑOZ et al., 2011) para a geração automatizada de metadados de proveniência durante a execução de experimentos de modelagem de nicho ecológico. Após a aquisição destes metadados, o openModeller irá utilizar os serviços de proveniência para o armazenamento destes metadados. O terceiro trabalho é o desenvolvimento de uma ferramenta para a análise dos metadados de proveniência utilizando técnicas de *machine learning* (RETTINGER et al., 2012). Os dois últimos trabalhos citados estão sendo desenvolvidos por alunos do Grupo de Estudos, Pesquisa e Extensão em Big Data, da Escola Politécnica da USP.

Além dos trabalhos em andamento, extensões desta pesquisa foram identificadas. Algumas propostas são:

- Definir novos mecanismos e estratégias para a descrição e aquisição de metadados de proveniência, com o objetivo de diminuir o esforço do pesquisador na descrição dos metadados e facilitar a sua operacionalização nas diversas etapas do ciclo de vida dos dados;
- Definir estratégias de *Visual Analytics* (BIKAKIS; SKOURLA; PAPASTEFANATOS, 2014) para a visualização e análise de extensos grafos de proveniência;
- Adaptar a estratégia computacional proposta para ambientes de computação de alto desempenho;
- Definir um componente de análise de dados para a arquitetura computacional de proveniência, com o objetivo de otimizar o processo de compreensão e validação de experimentos e recursos de dados;

- Estender o perfil de aplicação BioProv e a arquitetura computacional de proveniência para considerar a geração de um grafo do conhecimento da área de biodiversidade, que poderá relacionar dados taxonômicos, dados primários, imagens, vídeos, publicações, pesquisadores e instituições.

7 REFERÊNCIAS

AKABANE, G. K.; LOPES, C. P.; SILVA, F. P. DA. O Sistema de Rastreabilidade para a Sustentabilidade no Agronegócio Brasileiro. **Revista da Micro e Pequena Empresa**, v. 4, n. 2, p. 80–94, 13 abr. 2011.

ALCAMO, J.; BENNETT, E. M.; MILLENNIUM ECOSYSTEM ASSESSMENT (PROGRAM) (EDS.). **Ecosystems and human well-being: a framework for assessment**. Washington, DC: Island Press, 2003.

ALLARD, S. DataONE: Facilitating eScience through Collaboration. **Journal of eScience Librarianship**, p. 4–17, 2012.

ALTINTAS, I.; BARNEY, O.; JAEGER-FRANK, E. Provenance Collection Support in the Kepler Scientific Workflow System. In: MOREAU, L.; FOSTER, I. (Eds.). **Provenance and Annotation of Data**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 2006. p. 118–132.

ANTONIOU, G.; ANTONIOU, G. (EDS.). **A Semantic Web primer**. 3rd ed ed. Cambridge, Mass: MIT Press, 2012.

BACH, K. et al. A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. **Ecological Informatics**, Data platforms in integrative biodiversity research. v. 11, p. 16–24, Setembro 2012.

BAKER, E.; RYCROFT, S.; SMITH, V. Linking multiple biodiversity informatics platforms with Darwin Core Archives. **Biodiversity Data Journal**, v. 2, p. e1039, 8 jan. 2014.

BAKER, M. 1,500 scientists lift the lid on reproducibility. **Nature News**, v. 533, n. 7604, p. 452, 26 maio 2016.

BAKER, T.; COYLE, K. **Guidelines for Dublin Core Application Profiles**. Disponível em: <<http://dublincore.org/documents/profile-guidelines/>>.

BASS, L.; CLEMENTS, P.; KAZMAN, R. **Software Architecture in Practice**. 3 edition ed. Upper Saddle River, NJ: Addison-Wesley Professional, 2012.

BECKETT, D. et al. **RDF 1.1 Turtle**. Disponível em: <<https://www.w3.org/TR/turtle/>>.

BECKETT, D. **RDF 1.1 N-Triples**. Disponível em: <<https://www.w3.org/TR/n-triples/>>.

BELHAJJAME, K. et al. PROV-DM: The PROV Data Model. **W3C Recommendation**, 30 abr. 2013.

BERNERS-LEE, T. et al. The semantic web. **Scientific american**, v. 284, n. 5, p. 28–37, 2001.

BERNERS-LEE, T.; FIELDING, R. T.; MASINTER, L. **Uniform Resource Identifier (URI): Generic Syntax. RFC 3986 (Standard)**, jan. 2005. Disponível em: <<http://www.ietf.org/rfc/rfc3986.txt>>

BIKAKIS, N.; SKOURLA, M.; PAPASTEFANATOS, G. **rdf:SynopsViz – A Framework for Hierarchical Linked Data Visual Exploration and Analysis**. The Semantic Web: ESWC 2014 Satellite Events. **Anais...** In: EUROPEAN SEMANTIC WEB CONFERENCE. Springer, Cham, 25 maio 2014.

BLANKMAN, D.; MCGANN, J. **Ecological Metadata Language: Practical Application for Scientists**. Albuquerque, NM: LTER Network Office, 2003..

BOAKES, E. H. et al. Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. **PLoS Biol**, v. 8, n. 6, p. e1000385, 1 jun. 2010.

BOOTH, T. H. et al. bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. **Diversity and Distributions**, v. 20, n. 1, p. 1–9, 1 jan. 2014.

BORGMAN, C. L. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology**, v. 63, n. 6, p. 1059–1078, 2012.

BRICKLEY, D.; GUHA, R. V.; MCBRIDE, B. **RDF Schema 1.1. W3C Recommendation**, 25 fev. 2014.

BRICKLEY, D.; MILLER, L. **FOAF Vocabulary Specification**. Disponível em: <<http://xmlns.com/foaf/spec/>>.

BRINEY, K. **Data Management for Researchers: Organize, maintain and share your data for research success**. [s.l.] Pelagic Publishing, 2015.

BUITRAGO, L. et al. Plinian Core Profiles. Facilitating the documentation and access to biological information of exotic and invasive species. **reponame: Repositorio Institucional de Documentación Científica Humboldt Digital**, 5 dez. 2016.

BUNEMAN, P.; KHANNA, S.; TAN, W.-C. **Data provenance: Some basic issues**. International Conference on Foundations of Software Technology and Theoretical Computer Science. **Anais...** Springer, 2000 Disponível em: <http://link.springer.com/chapter/10.1007/3-540-44450-5_6>.

BUTLER, D. Mashups mix data into global service. **Nature**, v. 439, n. 7072, p. 6–7, 5 jan. 2006.

CALLAHAN, S. P. et al. **VisTrails: Visualization Meets Data Management**. Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. **Anais...** SIGMOD '06. New York, NY, USA: ACM, 2006.

CANDELA, L. et al. An infrastructure-oriented approach for supporting biodiversity research. **Ecological Informatics**, 2014.

CANHOS, D. A. L.; SARAIVA, A. M. Sistemas de informação e ferramentas computacionais para pesquisa, educação e disseminação do conhecimento sobre polinizadores. In: **Polinizadores do Brasil - contribuição e perspectivas para a biodiversidade, uso sustentável, conservação e serviços ambientais**. São Paulo: EDUSP, 2012.

CANHOS, V. P. et al. Global Biodiversity Informatics: setting the scene for a “new world” of ecological forecasting. **Biodiversity Informatics**, v. 1, n. 0, 11 nov. 2004.

CASSEY, P.; BLACKBURN, T. M. Reproducibility and repeatability in ecology. **BioScience**, v. 56, n. 12, p. 958–959, 2006.

CHAMBERLAIN, S. et al. rgbif: Interface to the Global Biodiversity Information Facility API. **R package version 0.7**, v. 7, 2014.

CHAPMAN, A. D. Uses of primary species-occurrence data. 2005a.

CHAPMAN, A. D. **Principles of Data Quality**. Copenhagen: GBIF, 2005b.

CICCARESE, P. et al. PAV ontology: provenance, authoring and versioning. **Journal of Biomedical Semantics**, v. 4, n. 1, p. 37, 22 nov. 2013.

CORRÊA, P. et al. Computational Techniques for Biologic Species Distribution Modeling. <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-61692-871-1.ch015>, p. 308–325, 2011.

CORRÊA, P. L. P. **Arquitetura para Integração de Sistemas de Informação de Biodiversidade**: Assessoria do Ministério do Meio Ambiente na Integração de Sistemas de Informação em Biodiversidade no âmbito do PROBIO II. [s.l.] Ministério do Meio Ambiente - MMA, maio 2013.

COSTA, F. et al. **Capturing and Querying Workflow Runtime Provenance with PROV: A Practical Approach**. Proceedings of the Joint EDBT/ICDT 2013 Workshops. **Anais...**: EDBT '13. New York, NY, USA: ACM, 2013.

COSTELLO, M. J. et al. Biodiversity data should be published, cited, and peer reviewed. **Trends in Ecology & Evolution**, v. 28, n. 8, p. 454–461, Agosto 2013.

COSTELLO, M. J.; WIECZOREK, J. Best practice for biodiversity data management and publication. **Biological Conservation**, v. 173, p. 68–73, maio 2014.

CURBERA, F. et al. Business Provenance – A Technology to Increase Traceability of End-to-End Operations. In: MEERSMAN, R.; TARI, Z. (Eds.). **On the Move to Meaningful Internet Systems: OTM 2008**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 2008. p. 100–119.

CYGANIAK, R.; WOOD, D.; LANTHALER, M. RDF 1.1 Concepts and Abstract Syntax. **W3C Recommendation**, 25 fev. 2014.

DA CRUZ, S. M. S.; CAMPOS, M. L. M.; MATTOSO, M. **A Foundational Ontology to Support Scientific Experiments**. ONTOBRAS-MOST. **Anais...Citeseer**, 2012.

DA SILVA, D. L.; BATISTA, A.; CORRÊA, P. L. P. **Data Provenance in Environmental Monitoring**. In: THE 13TH IEEE INTERNATIONAL CONFERENCE ON MOBILE AD HOC AND SENSOR SYSTEMS (IEEE MASS 2016). Brasília: 2016

DA SILVA, D. L.; CORREA, P. L. P.; STANZANI, S. L. Gestão de Dados. In: **Gestão de Dados de Biodiversidade**. Ciclo de Capacitação em Monitoramento da Biodiversidade. 1. ed. Brasília: Instituto Chico Mendes de Conservação da Biodiversidade, 2015. p. 29–34.

DAMA INTERNATIONAL. **The DAMA Guide to the Data Management Body of Knowledge**. First edition ed. Bradley Beach, N.J.: Technics Publications, LLC, 2009.

DAVIDSON, S. B.; FREIRE, J. **Provenance and scientific workflows: challenges and opportunities**. Proceedings of the 2008 ACM SIGMOD international conference on Management of data. **Anais...ACM**, 2008.

DAVISON, A. et al. Sumatra: a toolkit for reproducible research. **Implementing reproducible research**, p. 57–79, 2014.

DE MARCO JÚNIOR, P.; SIQUEIRA, M. F. Como determinar a distribuição potencial de espécies sob uma abordagem conservacionista. **Megadiversidade**, v. 5, p. 65–76, 2009.

DEELMAN, E. et al. Metadata and provenance management. **Scientific Data Management: Challenges, Technology, and Deployment**, n. A. Shoshani and D. Rotem, 14 maio 2010.

DHAR, V. Data science and prediction. **Communications of the ACM**, v. 56, n. 12, p. 64–73, 1 dez. 2013.

DICKINSON, J. L.; ZUCKERBERG, B.; BONTER, D. N. Citizen Science as an Ecological Research Tool: Challenges and Benefits. **Annual Review of Ecology, Evolution, and Systematics**, v. 41, n. 1, p. 149–172, dez. 2010.

DONOHO, D. L. et al. Reproducible research in computational harmonic analysis. **Computing in Science & Engineering**, v. 11, n. 1, 2009.

DOU, L. et al. Kurator: A Kepler Package for Data Curation Workflows. **Procedia Computer Science**, v. 9, p. 1614–1619, 2012.

DUBLIN CORE METADATA INITIATIVE. **DCMI Metadata Terms**. Disponível em: <<http://dublincore.org/documents/dcmi-terms/>>.

DUCHARME, B. **Learning SPARQL: querying and updating with SPARQL 1.1**. Second edition ed. Sebastopol, CA: O'Reilly Media, 2013.

DUMA, D. et al. **Applying core scientific concepts to context-based citation recommendation**. Proceedings of the 10th edition of the Language Resources and Evaluation Conference. **Anais...**2016.

DUVAL, E. et al. Metadata Principles and Practicalities. **D-Lib Magazine**, v. 8, n. 4, abr. 2002.

ELITH, J.; LEATHWICK, J. R. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. **Annual Review of Ecology, Evolution, and Systematics**, v. 40, n. 1, p. 677–697, dez. 2009.

ENKE, N. et al. The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —. **Ecological Informatics**, v. 11, p. 25–33, Setembro 2012.

EYNDEN, V. VAN DEN; UK DATA ARCHIVE. **Managing and sharing data: best practice for researchers**. Colchester: UK Data Archive, 2011.

FEGRAUS, 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. **The Bulletin of the Ecological Society of America**, v. 86, n. 3, p. 158–168, 1 jul. 2005.

FEIGENBAUM, L. et al. **SPARQL 1.1 Protocol**. W3C Recommendation. Disponível em: <<https://www.w3.org/TR/sparql11-protocol/>>.

FERGUSON, A. R. et al. Big data from small data: data-sharing in the “long tail” of neuroscience. **Nature Neuroscience**, v. 17, n. 11, p. 1442–1447, 28 out. 2014.

FIELDING, R. T. **Architectural styles and the design of network-based software architectures**. [s.l.] University of California, Irvine, 2000.

FORZZA, R. C. et al. HERBÁRIO VIRTUAL REFLORA. **Unisanta BioScience**, v. 4, n. 7, p. 88–94, 4 jan. 2016.

FOWLER, M. **UML Distilled: A Brief Guide to the Standard Object Modeling Language**. Edição: 3 ed. [s.l.] Addison-Wesley Professional, 2003.

FREIRE, J. et al. Provenance for computational tasks: A survey. **Computing in Science & Engineering**, v. 10, n. 3, p. 11–21, 2008.

FREIRE, J. et al. Reproducibility using vistrails. In: **Implementing Reproducible Research**. [s.l.: s.n.]. p. 33–52.

FREW, J.; SLAUGHTER, P. **ES3: A Demonstration of Transparent Provenance for Scientific Computation**. Provenance and Annotation of Data and Processes. **Anais...** In: INTERNATIONAL PROVENANCE AND ANNOTATION WORKSHOP. Springer, Berlin, Heidelberg, 17 jun. 2008.

GADELHA, L. et al. **SiBBR: Uma Infraestrutura para Coleta, Integração e Análise de Dados sobre a Biodiversidade Brasileira**. VIII Brazilian e-Science Workshop (BRESCI 2014). **Anais...**2014.

GANDON, F.; SCHREIBER, G.; BECKETT, D. **RDF 1.1 XML Syntax**. Disponível em: <<https://www.w3.org/TR/rdf-syntax-grammar/>>.

GILLILAND, A. J. Setting the stage. **Introduction to metadata**, n. 2, p. 1–19, 2008.

GOBLE, C. A. et al. myExperiment: a repository and social network for the sharing of bioinformatics workflows. **Nucleic Acids Research**, v. 38, n. suppl 2, p. W677–W682, 1 jul. 2010.

GREENBERG, J. Understanding metadata and metadata schemes. **Cataloging & classification quarterly**, v. 40, n. 3–4, p. 17–36, 2005.

GROTH, P. et al. Requirements for Provenance on the Web. **International Journal of Digital Curation**, v. 7, n. 1, p. 39–56, 12 mar. 2012.

GROTH, P.; MILES, S.; MOREAU, L. **PReServ: Provenance recording for services**. Proceedings of the UK OST e-Science second All Hands Meeting 2005 (AHM'05). **Anais...**Nottingham,UK: 2005

GROTH, P.; MILES, S.; MOREAU, L. A model of process documentation to determine provenance in mash-ups. **ACM Transactions on Internet Technology**, v. 9, n. 1, p. 1–31, 1 fev. 2009.

GROTH, P.; MOREAU, L. PROV-Overview: An Overview of the PROV Family of Documents. **W3C Working Group Note NOTE-provoverview- 20130430**, abr. 2013.

GURALNICK, R.; HILL, A. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. **Bioinformatics (Oxford, England)**, v. 25, n. 4, p. 421–428, 15 fev. 2009.

GURALNICK, R. P. et al. Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. **ZooKeys**, n. 494, p. 133–154, 6 abr. 2015.

HARDISTY, A.; ROBERTS, D. A decadal view of biodiversity informatics: challenges and priorities. **BMC Ecology**, v. 13, n. 1, p. 1–23, 1 dez. 2013.

HARRINGTON, I. S. G. K. V. S. A. Examples of ecological data synthesis driven by rich metadata, and practical guidelines to use the Ecological Metadata Language specification to this end. **International Journal of Metadata, Semantics and Ontologies**, v. 6, n. 1, p. 46–55, 2011.

HARTIG, O. Provenance Information in the Web of Data. **LDOW**, v. 538, 2009.

HASLHOFER, B.; KLAS, W. A Survey of Techniques for Achieving Metadata Interoperability. **ACM Comput. Surv.**, v. 42, n. 2, p. 7:1–7:37, mar. 2010.

HEDDEN, H. Taxonomies and controlled vocabularies best practices for metadata. **Journal of Digital Asset Management**, v. 6, n. 5, p. 279–284, Outubro 2010.

HEIDORN, P. B. Shedding light on the dark data in the long tail of science. **Library Trends**, v. 57, n. 2, p. 280–299, 2008.

HEIDORN, P. B. Biodiversity informatics. **Bulletin of the American Society for Information Science and Technology**, v. 37, n. 6, p. 38–44, 2011.

HEY, T.; TANSLEY, S.; TOLLE, K. (EDS.). **The Fourth Paradigm: Data-Intensive Scientific Discovery**. 1 edition ed. Redmond, Wash.: Microsoft Research, 2009.

HIJMANS, R. J. et al. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v. 25, n. 15, p. 1965–1978, 1 dez. 2005.

HIJMANS, R. J. et al. Package “dismo”. **Circles**, v. 9, p. 1, 2016.

HIJMANS, R. J.; GRAHAM, C. H. The ability of climate envelope models to predict the effect of climate change on species distributions. **Global change biology**, v. 12, n. 12, p. 2272–2281, 2006.

HIRZEL, A. H.; LE LAY, G. Habitat suitability modelling and niche theory. **Journal of Applied Ecology**, v. 45, n. 5, p. 1372–1381, out. 2008.

HOLETSCHEK, J. et al. The ABCD of primary biodiversity data access. **Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology**, v. 146, n. 4, p. 771–779, 16 out. 2012.

HOWE, B. et al. **Database-as-a-service for long-tail science**. Scientific and Statistical Database Management. **Anais...** Springer, 2011 Disponível em: <http://link.springer.com/chapter/10.1007/978-3-642-22351-8_31>.

HUYNH, T. D.; MOREAU, L. ProvStore: A public provenance repository. In: **Provenance and Annotation of Data and Processes**. [s.l.] Springer, 2014. p. 275–277.

INTERNATIONAL ORGANIZATION OF STANDARDIZATION. **ISO 19115 - Geographic information — Metadata**. Suíça: International Organization of Standardization, 5 jan. 2003.

ISO/IEC 10746-1:1998. **Open Distributed Processing - Reference model: Overview** International Organization for Standardization, 1998.

ISO/IEC 10746-2:2009. **Open distributed processing - Reference model: Foundations** International Organization for Standardization, 2009.

ISO/IEC 10746-3:2009. **Open distributed processing - Reference model: Architecture** International Organization for Standardization, 2009. Disponível em: <http://www.iso.org/iso/catalogue_detail.htm?csnumber=55724>.

ISO/IEC 10746-4:1998. **Open Distributed Processing - Reference Model: Architectural semantics** International Organization for Standardization, 1998. Disponível em: <http://www.iso.org/iso/catalogue_detail.htm?csnumber=20698>.

JAFFE, M. **ProvenanceGroove Art Online (The Dictionary of Art)**. [s.l.: s.n.].

JAROENSUTASINEE, K.; PHEERA, W.; JAROENSUTASINEE, M. Online weather data analysis and visualization tools for applications in ecoinformatics. **Earth Science Informatics**, v. 7, n. 3, p. 205–213, 1 set. 2014.

JEBBAR, M.; SEKKAKI, A.; BENAMAR, O. **Integration of SOA and Cloud Computing in RM-ODP**. Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on. **Anais...IEEE**, 2012.

JOHNSON, N. F. Biodiversity Informatics. **Annual Review of Entomology**, v. 52, n. 1, p. 421–438, 2007.

JONES, A. C. et al. Evolution of the Catalogue of Life Architecture. In: SETCHI, R. et al. (Eds.). **Knowledge-Based and Intelligent Information and Engineering Systems**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 2010. p. 485–496.

JONES, A. C.; WHITE, R. J.; ORME, E. R. Identifying and relating biological concepts in the Catalogue of Life. **Journal of Biomedical Semantics**, v. 2, n. 1, p. 7, 17 out. 2011.

KONING, D.; SARKAR, I. N.; MORITZ, T. TaxonGrab: Extracting Taxonomic Names From Text. **Biodiversity Informatics**, v. 2, n. 0, 16 nov. 2005.

KUNZE, J. **Towards electronic persistence using ARK identifiers**. Proceedings of the 3rd ECDL Workshop on Web Archives. **Anais...2003**.

KUNZE, J. A.; BAKER, T. **RFC 5013 - The Dublin Core Metadata Element Set** The IETF Trust, , 2007. Disponível em: <<http://www.ietf.org/rfc/rfc5013.txt>>.

KWASNIKOWSKA, N.; MOREAU, L.; BUSSCHE, J. V. D. A Formal Account of the Open Provenance Model. **ACM Trans. Web**, v. 9, n. 2, p. 10:1–10:44, maio 2015.

LANTHALER, M.; GÜTL, C. **On using JSON-LD to create evolvable RESTful services**. Proceedings of the Third International Workshop on RESTful Design. **Anais...ACM**, 2012.

LEBO, T. et al. PROV-O: The PROV Ontology. **W3C Recommendation**, 30 abr. 2013.

LEVEQUE, R. J. Python tools for reproducible research on hyperbolic problems. **Computing in Science and Engineering**, v. 11, n. 1, p. 19–27, 2009.

LIU, L.; ÖZSU, M. T. (EDS.). **Encyclopedia of Database Systems**. 2009 edition ed. New York: Springer, 2009.

LYNCH, C. Big data: How do your data grow? **Nature**, v. 455, n. 7209, p. 28–29, Setembro 2008.

MALTA, M. C.; BAPTISTA, A. A. **A method for the development of dublin core application profiles (Me4DCAP V0. 2): detailed description**. Proceedings of the 2013 International Conference on Dublin Core and Metadata Applications. **Anais...Citeseer**, 2013.

MALTA, M. C.; BAPTISTA, A. A. A panoramic view on metadata application profiles of the last decade. **International Journal of Metadata, Semantics and Ontologies**, v. 9, n. 1, p. 58, 2014.

MCPHILLIPS, T. et al. YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. **arXiv:1502.02403 [cs]**, 9 fev. 2015.

MESIROV, J. P. COMPUTER SCIENCE. Accessible Reproducible Research. **Science (New York, N.Y.)**, v. 327, n. 5964, 22 jan. 2010.

METADATA AD HOC WORKING GROUP. Content Standard for Digital Geospatial Metadata. **Federal Geographic Data Committee, Washington DC FGDC-STD-001-1998**, 1998.

MICHENER, W. et al. Nongeospatial metadata for the ecological sciences. **Ecological Applications**, v. 7, n. 1, p. 342, 1997.

MICHENER, W. K. et al. Long term ecological research and information management. **Ecological Informatics**, Special Issue: 5th Anniversary. v. 6, n. 1, p. 13–24, jan. 2011.

MICHENER, W. K. et al. Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. **Ecological Informatics**, v. 11, p. 5–15, set. 2012.

MICHENER, W. K.; JONES, M. B. Ecoinformatics: supporting ecology as a data-intensive science. **Trends in Ecology & Evolution**, Ecological and evolutionary informatics. v. 27, n. 2, p. 85–93, fev. 2012.

MISSIER, P. et al. Taverna, Reloaded. In: GERTZ, M.; LUDÄSCHER, B. (Eds.). **Scientific and Statistical Database Management**. Lecture Notes in Computer Science. [s.l.] Springer Berlin Heidelberg, 2010. p. 471–481.

MISSIER, P. et al. **D-PROV: extending the PROV provenance model with workflow structure**. TaPP. **Anais...2013**.

MISSIER, P.; BELHAJJAME, K.; CHENEY, J. **The W3C PROV family of specifications for modelling provenance metadata**. Proceedings of the 16th International Conference on Extending Database Technology. **Anais...ACM**, 2013.

MOREAU, L. et al. Special Issue: The First Provenance Challenge. **Concurrency and Computation: Practice and Experience**, v. 20, n. 5, p. 409–418, 10 abr. 2008.

MOREAU, L. et al. The Open Provenance Model core specification (v1.1). **Future Generation Computer Systems**, v. 27, n. 6, p. 743–756, jun. 2011.

MOREAU, L. et al. PROV-AQ: Provenance Access and Query. **W3C Working Group Note**, 30 abr. 2013.

MOREAU, L. et al. The rationale of PROV. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 35, p. 235–257, dez. 2015.

MOREAU, L.; GROTH, P. **Provenance: An Introduction to PROV**. [s.l.] Morgan & Claypool Publishers, 2013.

MORRIS, R. A. et al. Discovery and Publishing of Primary Biodiversity Data associated with Multimedia Resources: The Audubon Core Strategies and Approaches. **Biodiversity Informatics**, v. 8, n. 2, 9 jul. 2013.

MUNISWAMY-REDDY, K.-K. et al. **Provenance-Aware Storage Systems**. USENIX Annual Technical Conference, General Track. **Anais...**2006.

MUÑOZ, M. E. DE S. et al. openModeller: a generic approach to species' potential distribution modelling. **Geoinformatica**, v. 15, n. 1, p. 111–135, 1 jan. 2011.

MURTA, L. et al. noworkflow: Capturing and analyzing provenance of scripts. In: **Provenance and Annotation of Data and Processes**. [s.l.] Springer, 2014. p. 71–83.

NILSSON, M.; BAKER, T.; JOHNSTON, P. **The Singapore Framework for Dublin Core Application Profiles**. Disponível em: <<http://dublincore.org/documents/singapore-framework/>>.

NISO, N. I. S. O. (U. S. . (ED.). **Understanding metadata**. Bethesda, MD: NISO Press, 2004.

NOWACK, B. **The Semantic Web Technology Stack)not a piece of cake...)** Linked Data Developer, , 2009.

PAGE, R. D. M. Taxonomic names, metadata, and the Semantic Web. **Biodiversity Informatics**, v. 3, n. 0, 2 jun. 2006.

PAGE, R. D. M. Biodiversity informatics: the challenge of linking data and the role of shared identifiers. **Briefings in Bioinformatics**, v. 9, n. 5, p. 345–354, 1 set. 2008.

PALMER, C. L. et al. **Data curation for the long tail of science: The case of environmental sciences**. Third International Digital Curation Conference, Washington, DC. **Anais...**2007.

PANCERELLA, C.; MYERS, J.; RAHN, L. **Data Provenance in the Collaboratory for Multiscale Chemical Science (CMCS)**. . In: WORKSHOP ON DATA DERIVATION AND PROVENANCE. Chicago, Illinois: 2002.

PENG, R. D. Reproducible research and Biostatistics. **Biostatistics**, v. 10, n. 3, p. 405–408, 1 jul. 2009.

PENG, R. D. Reproducible Research in Computational Science. **Science**, v. 334, n. 6060, p. 1226–1227, 2 dez. 2011.

PENG, R. D.; DOMINICI, F.; ZEGER, S. L. Reproducible Epidemiologic Research. **American Journal of Epidemiology**, v. 163, n. 9, p. 783–789, 1 maio 2006.

PENNINGTON, D. D. et al. Ecological Niche Modeling Using the Kepler Workflow System. In: TAYLOR, I. J. et al. (Eds.). . **Workflows for e-Science**. [s.l.] Springer London, 2007. p. 91–108.

PETERSON, A. T. et al. The big questions for biodiversity informatics. **Systematics and Biodiversity**, v. 8, n. 2, p. 159–168, 23 jun. 2010.

PETERSON, A. T.; SOBERÓN, J.; KRISHTALKA, L. A global perspective on decadal challenges and priorities in biodiversity informatics. **BMC Ecology**, v. 15, p. 15, 2015.

PHILIP CHEN, C. L.; ZHANG, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. **Information Sciences**, v. 275, p. 314–347, 10 ago. 2014.

PHILLIPS, S. J. et al. Opening the black box: an open-source release of Maxent. **Ecography**, p. n/a-n/a, 1 mar. 2017.

PHILLIPS, S. J.; ANDERSON, R. P.; SCHAPIRE, R. E. Maximum entropy modeling of species geographic distributions. **Ecological Modelling**, v. 190, n. 3–4, p. 231–259, 25 jan. 2006.

PIMENTEL, J. F. N. et al. **Collecting and Analyzing Provenance on Interactive Notebooks: When IPython Meets noWorkflow**. 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP 15). **Anais...**2015.

POWELL, A. et al. **DCMI Abstract Model**. Disponível em: <<http://dublincore.org/documents/abstract-model/>>.

PROVOST, F.; FAWCETT, T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, v. 1, n. 1, p. 51–59, mar. 2013.

PRUD'HOMMEAUX, E.; BUIL-ARANDA, C. **SPARQL 1.1 Federated Query**. Disponível em: <<https://www.w3.org/TR/sparql11-federated-query/>>.

RAVEN, P. H.; BERLIN, B.; BREEDLOVE, D. E. The origins of taxonomy. **Science**, v. 174, n. 4015, p. 1210–1213, 1971.

RDC TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2006.

RETTINGER, A. et al. Mining the Semantic Web. **Data Min. Knowl. Discov.**, v. 24, n. 3, p. 613–662, maio 2012.

ROBERTSON, T. et al. The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. **PLoS ONE**, v. 9, n. 8, p. e102623, 6 ago. 2014.

RÜEGG, J. et al. Completing the data life cycle: using information management in macrosystems ecology research. **Frontiers in Ecology and the Environment**, v. 12, n. 1, p. 24–30, fev. 2014.

SAHOO, S.; SHETH, A. Provenir Ontology: Towards a Framework for eScience Provenance Management. **Kno.e.sis Publications**, 15 out. 2009.

SANTANA, F. S. et al. A reference business process for ecological niche modelling. **Ecological Informatics**, v. 3, n. 1, p. 75–86, 1 jan. 2008.

SCHALK, P. H. Management of marine natural resources through by biodiversity informatics. **Marine Policy**, v. 22, n. 3, p. 269–280, maio 1998.

SILVA, D. L. et al. **Diretrizes para a Integração de Dados de Biodiversidade**. 1. ed. Brasília: Ministério do Meio Ambiente, 2015.

SILVA, D. L. et al. **A Provenance-Based Approach for Reusing Biodiversity Data Products**. . In: THE 15TH INTERNATIONAL CONFERENCE ON INFORMATION & KNOWLEDGE ENGINEERING (IKE'16). Las Vegas: 2016

SILVA, D. L.; BATISTA, A.; CORRÊA, P. L. P. **Data Provenance in Environmental Monitoring**. 2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS). **Anais...** In: 2016 IEEE 13TH INTERNATIONAL CONFERENCE ON MOBILE AD HOC AND SENSOR SYSTEMS (MASS). out. 2016

SIMMHAN, Y. L.; PLAILE, B.; GANNON, D. A survey of data provenance in e-science. **ACM Sigmod Record**, v. 34, n. 3, p. 31–36, 2005.

SIMONS, N. Implementing DOIs for Research Data. **D-Lib Magazine**, v. 18, n. 5, p. 1–, 2012.

SMITH, A. B.; ROWE, K. C. **Testing methods for predicting mammalian species' responses to 20th century climate change in California**. [s.l.] California Institute for Energy and Environment, 2011. Disponível em: <<http://www.uc-ciee.org/additional-documents/8/521/121/nested>>.

SMITH, M. et al. **OWL 2 Web Ontology Language Document Overview (Second Edition)**. Disponível em: <<https://www.w3.org/TR/owl2-overview/>>.

SOBERÓN, J. Grinnellian and Eltonian niches and geographic distributions of species. **Ecology Letters**, v. 10, n. 12, p. 1115–1123, 1 dez. 2007.

SOBERÓN, J.; NAKAMURA, M. Niches and distributional areas: Concepts, methods, and assumptions. **Proceedings of the National Academy of Sciences**, v. 106, n. Supplement 2, p. 19644–19650, 17 nov. 2009.

SOBERÓN, J.; PETERSON, T. Biodiversity informatics: managing and applying primary biodiversity data. **Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences**, v. 359, n. 1444, p. 689–698, 29 abr. 2004.

SOLDATOVA, L. N.; KING, R. D. An ontology of scientific experiments. **Journal of the Royal Society Interface**, v. 3, n. 11, p. 795–803, 22 dez. 2006.

SPORNY, M. et al. **JSON-LD 1.0: A JSON-based Serialization for Linked Data**. Disponível em: <<https://www.w3.org/TR/json-ld/>>.

STEIN, B. R.; WIECZOREK, J. R. Mammals of the World: MaNIS as an example of data integration in a distributed network environment. **Biodiversity Informatics**, v. 1, n. 0, 11 nov. 2004.

STODDEN, V.; LEISCH, F.; PENG, R. D. (EDS.). **Implementing Reproducible Research**. Boca Raton: Chapman and Hall/CRC, 2014.

STRASSER, C. et al. Primer on Data Management: What you always wanted to know. 2012.

SULLIVAN, B. L. et al. eBird: A citizen-based bird observation network in the biological sciences. **Biological Conservation**, v. 142, n. 10, p. 2282–2292, Outubro 2009.

TAN, W. C. Provenance in Databases: Past, Current, and Future. **IEEE Data Eng. Bull.**, v. 30, n. 4, p. 3–12, 2007.

TDWG. **Darwin Core Terms: A quick reference guide**. Disponível em: <<http://rs.tdwg.org/dwc/terms/index.htm>>.

THE W3C SPARQL WORKING GROUP. **SPARQL 1.1 Overview**. Disponível em: <<https://www.w3.org/TR/sparql11-overview/>>.

TOWNSEND PETERSON, A.; PAPEŞ, M.; EATON, M. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. **Ecography**, v. 30, n. 4, p. 550–560, 1 ago. 2007.

VAN DE SOMPEL, H. et al. Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. **International Journal of Digital Curation**, v. 9, n. 1, p. 331–342, 17 jun. 2014.

VANDEWALLE, P.; KOVACEVIC, J.; VETTERLI, M. Reproducible research in signal processing. **IEEE Signal Processing Magazine**, v. 26, n. 3, p. 37–47, maio 2009.

VINES, T. H. et al. The Availability of Research Data Declines Rapidly with Article Age. **Current Biology**, v. 24, n. 1, p. 94–97, 6 jan. 2014.

WALLIS, J. C.; ROLANDO, E.; BORGMAN, C. L. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. **PLOS ONE**, v. 8, n. 7, p. e67332, 23 jul. 2013.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação**. [s.l.: s.n.].

WEIBEL, S. et al. **Dublin Core Metadata for Resource Discovery**. [s.l.] RFC Editor, set. 1998. Disponível em: <<https://www.rfc-editor.org/info/rfc2413>>.

WIECZOREK, J. et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. **PLoS ONE**, v. 7, n. 1, p. e29715, 6 jan. 2012.

WOLSTENCROFT, K. et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. **Nucleic Acids Research**, v. 41, n. W1, p. W557–W561, 1 jul. 2013.

WOODLEY, M. S.; CLEMENT, G.; WINN, P. **Dublin Core Glossary**. Disponível em: <<http://dublincore.org/documents/2001/04/12/usageguide/glossary.shtml>>.

YESSON, C. et al. How Global Is the Global Biodiversity Information Facility? **PLoS ONE**, v. 2, n. 11, p. e1124, 7 nov. 2007.

APÊNDICE A – Description Set Profile (DSP) do Perfil de Aplicação Proposto e Vocabulários controlados

Entidade	Experimento Científico	Classe	Bi PROV: Experiment (PROV: Entity)
Descrição	Propriedade	Abrangência	Textual
identificador	dcterms: identifier	-	sim
título	dcterms: title	-	sim
descrição	dcterms: description	-	sim
tipo da entidade	dcterms: type	-	sim
justificativa	bi PROV: cause	-	sim
delimitação do Problema	bi PROV: problemDelimitation	-	sim
hipóteses	bi PROV: hypothesis	-	sim
objetivos	bi PROV: goals	-	sim
metodologia	bi PROV: method	-	sim
requisitos	bi PROV: has requirements	-	sim
cobertura taxonômica	bi PROV: taxonomicCoverage	-	sim
cobertura temporal	dcterms: temporal	-	sim
cobertura espacial	dcterms: spatial	-	sim
data/hora de início	prov: startedAtTime	xsd: dateTime	não
data/hora de término	prov: endedAtTime	xsd: dateTime	não
responsável	prov: wasAttributedTo	prov: Agent	não
resultados	bi PROV: has results	-	sim
localização	prov: atLocation	-	sim
veja também	rdfs: seeAlso	-	não

Entidade	Projeto de Pesquisa	Classe	bioprov:ResearchProject (prov:Entity)
Descrição	Propriedade	Abrangência	Textual
identificador	dcterms:identifier	-	sim
título	dcterms:title	-	sim
descrição	dcterms:description	-	sim
tipo da entidade	dcterms:type	-	sim
área de estudo	bioprov:studyAreaDescription	-	sim
cobertura taxonômica	bioprov:taxonomicCoverage	-	sim
cobertura temporal	dcterms:temporal	-	sim
cobertura espacial	dcterms:spatial	-	sim
objetivos	bioprov:goals	-	sim
financiador	bioprov:funding	-	sim
responsável	prov:wasAttributedTo	prov:Agent	não
Resultados	bioprov:has_results	-	sim
localização	prov:atLocation	-	sim
veja também	rdfs:seeAlso	-	não
Entidade	Atividades	Classe	Prov:Activity
Descrição	Propriedade	Abrangência	Textual
identificador	dcterms:identifier	-	sim
título	dcterms:title	-	sim
descrição	dcterms:description	-	sim
data/hora de início	prov:startedAtTime	xsd:dateTime	sim
data/hora de término	prov:endedAtTime	xsd:dateTime	sim
data/hora de invalidação	prov:invalidatedAtTime	xsd:dateTime	não
duração	time:hasDuration	-	sim
tipo da atividade	bioprov:activityType	bioprov:BioActivityType	não

responsável	prov:wasAssociatedWith	prov:Agent	não
Insumos	prov:used	prov:Entity	não
Iniciado a partir de	prov:wasInformedBy	prov:Activity	não
decisões	bioprov:decisions	-	sim
premissas	bioprov:assumptions	-	sim
arquitetura computacional utilizada	bioprov:computationalArchitectureUsed	-	sim
dependências funcionais	bioprov:FunctionalDependencies	-	sim
localização	prov:atLocation	-	sim
veja também	rdfs:seeAlso	-	não
Entidade	Recurso de Dados	Classe	bioprov:DataResource (prov:Entity)
Descrição	Propriedade	Abrangência	Textual
identificador	dcterms:identifier	-	sim
título	dcterms:title	-	sim
descrição	dcterms:description	-	sim
tipo da entidade	dcterms:type	-	sim
formato do dado	bioprov:formatType	bioprov:BioFormatType	não
tipo do dado	bioprov:entityType	bioprov:BioEntityType	não
versão	dcterms:isVersionOf	-	não
responsável	prov:wasAttributedTo	prov:Agent	não
fonte dos dados	dcterms:source	-	sim
data de criação	dcterms:created	-	sim
gerado a partir de	prov:wasGeneratedBy	prov:Activity	não
derivado de	prov:wasDerivedFrom	prov:Entity	não
localização	prov:atLocation	-	sim
latitude	geo:lat	-	sim
longitude	geo:lon	-	sim

valor	rdf:value	-	sim
nome da espécie	dwc:scientificName	-	sim
possui membros	prov:hadMember	prov:Entity	não
licença de uso	dcterms:license	-	sim
direitos de uso	dcterms:right	-	sim
citação bibliográfica	dcterms:bibliographicCitation	-	sim
veja também	rdfs:seeAlso	-	não
o mesmo	owl:sameAs	-	não

Entidade	Recurso	Classe	bioprov:Resource (prov:Entity)
Descrição	Propriedade	Abrangência	Textual
identificador	dcterms:identifier	-	sim
título	dcterms:title	-	sim
descrição	dcterms:description	-	sim
tipo da entidade	dcterms:type	-	sim
formato do dado	bioprov:entityFormat	bioprov:Format	não
tipo do recurso	bioprov:entityType	bioprov:Type	não
responsável	prov:wasAttributedTo	prov:Agent	não
fonte do recurso	dcterms:source	-	sim
data de criação	dcterms:created	-	sim
valor	rdf:value	-	sim
gerados a partir de	prov:wasGeneratedBy	prov:Activity	não
derivado de	prov:wasDerivedFrom	prov:Entity	não
localização	prov:atLocation	-	sim
possui membros	prov:hadMember	prov:Entity	não
veja também	rdfs:seeAlso	-	não

o mesmo	owl:sameAs	-	não
---------	------------	---	-----

Entidade	Instituição	Classe	biprov:Organization (prov:Agent)
Descrição	Propriedade	Abrangência	Textual
identificador	dcterms:identifier	-	não
nome	foaf:name	-	sim
e-mail	foaf:mbox	-	sim
página Web	foaf:homepage	-	sim
descrição	dcterms:description	-	sim
localização	prov:atLocation	prov:Location, geo:SpatialThing	sim
representa	prov:actedOnBehalfOf	prov:Agent	não
veja também	rdfs:seeAlso	-	não

Entidade	Pesquisador	Classe	biprov:Researcher (prov:Agent)
Descrição	Propriedade	Abrangência	Textual
identificador	dcterms:identifier	-	não
nome	foaf:name	-	sim
e-mail	foaf:mbox	-	sim
página Web	foaf:homepage	-	sim
descrição	dcterms:description	-	sim
localização	prov:atLocation	prov:Location, geo:SpatialThing	sim
representa	prov:actedOnBehalfOf	prov:Agent	não
veja também	rdfs:seeAlso	-	não

Controlled Vocabularies		
bioprov:activityType	bioprov:Type	bioprov:Format
acquisition	DataResource	Dataset
Recovery	InputParameter	Record
Integration	Configuration	Checklist
Combination	Software	Catalog
Preparation	Equipment	DigitalFile
Processing	Service	RasterFile
Analysis		VectorFile
Publication		PhysicalFile
Sharing		SourceCode

APÊNDICE B – Repositórios de Códigos-Fonte dos Componentes do Protótipo Funcional

- **Servidor de Proveniência BioProv (BioProv Server):** Aplicação servidora que disponibiliza os serviços de proveniência para registro, consulta e visualização dos metadados de proveniência. Os demais componentes que compõem o servidor também estão implementados neste projeto.
 - <https://gitlab.com/daniellins/bioprov>
- **Pacote R do BioProv Client:** Pacote, da linguagem R, que disponibiliza funções para o registro e consulta de metadados de proveniência durante a execução de scripts. Este pacote se conecta ao servidor de proveniência BioProv por meio dos serviços de proveniência.
 - <https://gitlab.com/daniellins/bioprov-R-client>
- **Extensão (*plug-in*) do Chrome BioProv:** Extensão do navegador Chrome que consulta automaticamente a proveniência de recursos de dados publicados em páginas Web, seguindo as regras descritas no W3C PROV (PROV-AQ).
 - https://gitlab.com/daniellins/bioprov_browser_plugin

APÊNDICE C – Script do experimento realizado no Estudo de Caso 1.

```
#Author: Daniel Lins da Silva

#install.packages("raster")
#install.packages("rgdal")
#install.packages("dismo")
#install.packages("XML")
#install.packages("maps")
#install.packages("rgbif")
#install.packages("devtools")
#devtools::install_git("git@gitlab.com:daniellins/bioprov-R-client.git", branch = "ziziphus")

library(raster)
library(rgdal)
library(dismo)
library(XML)
library(maps)
library(BioProv)
library(jsonlite)
library(httr)
library(rgbif)

ag_daniel <- createAgentProvenance(uri = 'http://bioprov.poli.usp.br/agents/6613529',
                                   type = 'Researcher',
                                   name = 'Daniel Lins da Silva',
                                   identifier = 'http://bioprov.poli.usp.br/agents/6613529',
                                   mbox = 'mailto:daniellins@usp.br',
                                   seeAlso = 'http://lattes.cnpq.br/6502450550040226')

ag_andre <- createAgentProvenance(uri = 'http://bioprov.poli.usp.br/agents/8788438',
                                   type = 'Researcher',
                                   name = 'Andre Filipe de Moraes Batista',
                                   identifier = 'http://bioprov.poli.usp.br/agents/8788438',
                                   mbox = 'mailto:andrefmb@usp.br',
                                   seeAlso = 'http://lattes.cnpq.br/2015345575533420')
```

```

### Provenance Record - Bundle of Experiment
bundle <- createProvenanceBundle(uri = 'http://bioprov.poli.usp.br/bundle/2',
                                type = 'Experiment',
                                title = 'Habitat Adequability of Ziziphus juazeiro specie in the Brazilian Semi-
arid Region through Ecological Niche Modeling technique',
                                description = 'This work aims to show the relationship between physical aspects and
the occurrence of endemic species of the Brazilian semi-arid
region through ecological niche modeling technique aided by remote sensing products. It is an interdisciplinary
project, integrating Brazilian and Spanish research centers in the
study fields of biogeography, remote sensing, statistics and computer science. We use the R programming language and
DISMO Package to execute this experiment.',
                                wasAttributedTo = list(ag_daniel, ag_andre),
                                source = 'https://gitlab.com/daniellins/Prov_ENM_Ziziphus/',
                                date = date(),
                                creator = ag_daniel)

source('./SDM_R_functions.R')

occ <- gbifDataSearchingAndPreparing("Ziziphus joazeiro", country = "BR", hasCoordinate = TRUE, hasGeospatialIssue =
FALSE)

### Provenance Record - Agents
ag_gbif <- createAgentProvenance(uri = 'http://gbif.org/',
                                type = 'Institution',
                                name = 'Global Biodiversity Information Facility',
                                identifier = 'GBIF',
                                homepage = 'http://gbif.org/',
                                description = 'The Global Biodiversity Information Facility (GBIF) is an
international open data infrastructure.')

### Provenance Record - Input Data/ Input Parameters and Output Data (Entities)
e_rgbif <- createEntityProvenance(bundle = bundle,
                                identifier = "rgbif",
                                type = "Software",

```

```

uri = 'http://cran.rstudio.com/web/packages/rgbif',
title = 'rgbif',
description = 'Interface to the Global Biodiversity Information Facility API',
source = 'http://cran.rstudio.com/web/packages/rgbif/',
wasAttributedTo = "http://ropensci.org/")

e_gbifDataSearchingAndPreparing <- createEntityProvenance(bundle = bundle,
  identifier = "acquisitionAndPreparationFunction",
  type = "Software",
  uri =
'https://gitlab.com/daniellins/R_functions/blob/master/SDM_R_functions.R',
  title = 'GBIF data acquisition and preparation function',
  source =
'https://gitlab.com/daniellins/R_functions/blob/master/SDM_R_functions.R',
  wasAttributedTo = ag_daniel,
  wasDerivedFrom = e_rgbif)

e_param_species <- createEntityProvenance(bundle = bundle, uri = 'http://bioprov.poli.usp.br/param/species',
  identifier = 'species', type = 'InputParameter', value = 'Ziziphus joazeiro')
e_param_country <- createEntityProvenance(bundle = bundle, uri = 'http://bioprov.poli.usp.br/param/country',
  identifier = 'country', type = 'InputParameter', value = 'BR')
e_param_coordinate <- createEntityProvenance(bundle = bundle, uri =
'http://bioprov.poli.usp.br/param/hasCoordinate', identifier = 'hasCoordinate', type = "InputParameter", value =
'TRUE')
e_param_geospatial <- createEntityProvenance(bundle = bundle, uri =
'http://bioprov.poli.usp.br/param/hasGeospatialIssue', identifier = 'hasGeospatialIssue', type = 'InputParameter',
value = 'FALSE')

e_dataset <- createEntityProvenance(bundle = bundle,
  identifier = 'http://bioprov.poli.usp.br/occ_ziziphus_juazeiro',
  type = 'Dataset',
  uri = 'http://bioprov.poli.usp.br/occ_ziziphus_juazeiro',
  title = 'Occurrence Dataset of Ziziphus juazeiro in Brazil',
  wasAttributedTo = ag_gbif)

### Provenance Record - Data Preparing Activity (used Input Data/Parameters, generated Output Data)
act_dataSearching <- createActivityProvenance(bundle = bundle,

```

```

uri = 'http://bioprov.poli.usp.br/acquisition',
type = 'Acquisition',
title = 'Retrieving and Preparing Occurrence Records from GBIF',
used = list(e_gbifDataSearchingAndPreparing, e_param_species,
e_param_country, e_param_coordinate, e_param_geospatial),
generated = e_dataset)

map()
points(occ$Longitude, occ$Latitude, col="red", pch=19, xlab="Longitude", ylab="Latitude")
plot(occ$Longitude, occ$Latitude, xlim=c(-90,-25), ylim=c(-60,20), col="red", pch=19, xlab="Longitude", ylab="Latitude")
map(add=T)

files <- list.files("clima/", pattern='grd', full.names=TRUE)
files
predictors <- stack(files)
predictors
names(predictors)
plot(predictors, main=NA)

zizi.coord <- data.frame(occ$Longitude, occ$Latitude)
bio.ziziphus <- bioclim(predictors, zizi.coord)

### Provenance Record - Input Data/ Input Parameters and Output Data (Entities)
e_environmentalLayers <- createEntityProvenance(bundle = bundle,
uri = 'http://bioprov.poli.usp.br/bioclim_env_layers',
identifier = 'http://bioprov.poli.usp.br/bioclim_env_layers',
type = 'RasterLayer',
title = 'Worldclim Environmental Layers',
source =
'https://gitlab.com/daniellins/Prov_ENM_Ziziphus/tree/master/clima',
wasAttributedTo = 'http://www.worldclim.org/',
description = "Environmental Layers from WorldClim Website.")

```

```

e_enmAlgorithm <- createEntityProvenance(bundle = bundle,
                                         identifier = "Bioclim",
                                         type = "Software",
                                         uri = "http://cran.r-project.org/web/packages/dismo",
                                         title = "BIOCLIM implementation of DISMO package",
                                         hasVersion = '1.0-15',
                                         source = "http://cran.r-project.org/src/contrib/dismo_1.0-15.tar.gz",
                                         wasAttributedTo = "http://desp.ucdavis.edu/people/robert-j-hijmans/",
                                         seeAlso =
list('http://finzi.psych.upenn.edu/library/dismo/html/bioclim.html', 'http://cran.r-
project.org/web/packages/dismo/dismo.pdf'))

e_resultModeling <- createEntityProvenance(bundle = bundle,
                                           identifier = "ZiziphusModel",
                                           type = "Model",
                                           title = 'Species Distribution Model of the Ziziphus juazeiro',
                                           wasAttributedTo = ag_daniel)

### Provenance Record - ENM Activity
act_dataModeling <- createActivityProvenance(bundle = bundle,
                                             uri = 'http://bioprov.poli.usp.br/analysing',
                                             type = 'Analysis',
                                             title = 'Species Distribution Modeling with the Bioclim Algorithm',
                                             used = list(e_dataset, e_environmentalLayers, e_enmAlgorithm),
                                             generated = e_resultModeling)

plot(bio.ziziphus)
View(zizi.coord)

detach("package:stats")
map.ziziphus <- predict(bio.ziziphus, predictors)
plot(map.ziziphus)
map(add=T)

```

```

### Provenance Record - Input Data/ Input Parameters and Output Data (Entities)
e_MapresultModeling <- createEntityProvenance(bundle = bundle,
                                              identifier = "ZiziphusModelProjection",
                                              type = 'rasterLayer',
                                              title = 'Species Distribution Model Projection',
                                              description = 'Species Distribution Model projected on the map',
                                              wasAttributedTo = ag_daniel,
                                              wasDerivedFrom = e_resultModeling)

e_plotFunction <- createEntityProvenance(bundle = bundle,
                                         identifier = "Plot",
                                         type = "Software",
                                         uri = "http://cran.r-project.org/web/packages/raster#plot",
                                         title = "Plot function of the RASTER package from R Programming
Language",
                                         hasVersion = '2.5-2',
                                         source = "http://cran.r-project.org/src/contrib/raster_2.5-2.tar.gz",
                                         wasAttributedTo = "http://desp.ucdavis.edu/people/robert-j-hijmans",
                                         seeAlso = 'http://cran.r-project.org/web/packages/raster/raster.pdf')

### Provenance Record - Data Model Projecting Activity
act_dataPlotting <- createActivityProvenance(bundle = bundle,
                                              uri = 'http://bioprov.poli.usp.br/projecting',
                                              type = 'Projecting',
                                              title = 'Projecting model on a map',
                                              used = list(e_plotFunction, e_resultModeling),
                                              generated = e_MapresultModeling)

biozizi.rec<-reclassify(map.ziziphus,c(0,0.05,0.25, 0.051,0.1,0.5, 0.11,0.4,1))
plot(biozizi.rec)
map(add=T)

### Provenance Record - Input Data/ Input Parameters and Output Data (Entities)
e_FinalresultModeling <- createEntityProvenance(bundle = bundle,

```



```

uri = 'http://bioprov.poli.usp.br/models/ziziphusJuazeiro',
identifier = 'http://bioprov.poli.usp.br/models/ziziphusJuazeiro',
type = 'rasterLayer',
title = 'Map with the final result of SDM experiment',
source =
'https://gitlab.com/daniellins/R_functions/blob/master/ENM_result_ziziphus.jpg',
wasAttributedTo = list(ag_daniel, ag_andre),
wasDerivedFrom = e_MapresultModeling)

### Provenance Record - Map Improvement Activity
act_dataImprovement <- createActivityProvenance(bundle = bundle,
uri = 'http://bioprov.poli.usp.br/Improvement',
type = 'Improvement',
title = 'Map Improvement using Reclassify function of RASTER package',
seeAlso =
'https://www.rdocumentation.org/packages/raster/versions/2.5-2/topics/reclassify',
used = e_MapresultModeling,
generated = e_FinalresultModeling)

getProvenanceOfEntity(uri = e_FinalresultModeling, format = 'png')
getProvenanceOfEntity(uri = e_FinalresultModeling, format = 'rdf')

```

Link para acesso ao arquivo original: https://gitlab.com/daniellins/Prov_ENM_Ziziphus/blob/master/script_niche_modeling.R

APÊNDICE D – Relatório de Proveniência do Estudo de Caso 1 (formato PNG).

Link para acesso ao arquivo original: https://gitlab.com/daniellins/Prov_ENM_Ziziphus/tree/master/Results

APÊNDICE E – Trecho do Relatório de Proveniência do Estudo de Caso 1 (Formato RDF/Turtle).

```
<http://bioprov.poli.usp.br/models/ziziphusJuazeiro> a prov:Entity ;
  dct:created "2017-04-23 02:06:15" ;
  dct:identifier "http://bioprov.poli.usp.br/models/ziziphusJuazeiro" ;
  dct:source <https://gitlab.com/daniellins/R_functions/blob/master/ENM_result_ziziphus.jpg> ;
  dct:title "Map with the final result of SDM experiment" ;
  dct:type "rasterLayer" ;
  prov:wasAttributedTo <http://bioprov.poli.usp.br/agents/6613529>,
    <http://bioprov.poli.usp.br/agents/8788438> ;
  prov:wasDerivedFrom <http://bioprov.poli.usp.br/entities/90a9b45e-247a-4e62-8257-b0666c2b00e1> ;
  prov:wasGeneratedBy <http://bioprov.poli.usp.br/Improvement> .

<http://bioprov.poli.usp.br/Improvement> a prov:Activity ;
  dct:created "2017-04-23 02:06:15" ;
  dct:title "Map Improvement using Reclassify function of RASTER package" ;
  dct:type "Improvement" ;
  rdfs:seeAlso <https://www.rdocumentation.org/packages/raster/versions/2.5-2/topics/reclassify> ;
  prov:used <http://bioprov.poli.usp.br/entities/90a9b45e-247a-4e62-8257-b0666c2b00e1> .

<http://bioprov.poli.usp.br/acquisition> a prov:Activity ;
  dct:created "2017-04-23 02:06:12" ;
  dct:title "Retrieving and Preparing Occurrence Records from GBIF" ;
  dct:type "Acquisition" ;
  prov:used <http://bioprov.poli.usp.br/param/country>,
    <http://bioprov.poli.usp.br/param/hasCoordinate>,
    <http://bioprov.poli.usp.br/param/hasGeospatialIssue>,
    <http://bioprov.poli.usp.br/param/species>,
    <https://gitlab.com/daniellins/R_functions/blob/master/SDM_R_functions.R> .
```

Link para acesso ao arquivo original: https://gitlab.com/daniellins/Prov_ENM_Ziziphus/tree/master/Results

APÊNDICE F – Script do experimento realizado no Estudo de Caso 2.

```
#Author Daniel Lins da Silva

#install.packages(c('raster', 'dismo')) # packages created by Robert Hijmans for handling rasters and species
distribution modeling
#devtools::install_git('https://gitlab.com/daniellins/bioprov-R-client.git') #package created by Daniel Lins for
recording and retrieving provenance metadata
library(raster)
library(dismo)
library(rJava)
library(rgdal)
library(BioProv)
library(jsonlite)
library(httr)
library(XML)
library(maps)

### PROVENANCE - INITIAL METADATA (RESEARCH PROJECT AND EXPERIMENT DOCUMENTATION)

ag_smith <- createAgentProvenance(uri = 'http://orcid.org/0000-0002-6420-1659',
                                name = 'Adam B. Smith',
                                identifier = 'Smith',
                                mbox = 'mailto:adam@earthskysea.org',
                                seeAlso = 'http://www.earthskysea.org/')

research_project <-createEntityProvenance(uri = 'http://bioprov.poli.usp.br/projects/5', title = 'Testing methods
for predicting mammalian species responses to 20th century climate change in California',
                                entityType='ResearchProject', studyAreaDescription='Biodiversity',
taxonomicCoverage = 'Mammalian species', spatial='Western United States',
                                wasAttributedTo = ag_smith)

e_projectReport <-createEntityProvenance(uri = 'http://www.uc-ciee.org/Grinnellresurvey', title = 'Testing methods
for predicting mammalian species responses to 20th century climate change in California',
```

```

                                entityType='Publication', source='http://www.uc-
ciece.org/downloads/Grinnellresurvey.Smith.pdf', created = '2011-03-31',
                                wasDerivedFrom=research_project)

ag_daniel <- createAgentProvenance(uri = 'http://orcid.org/0000-0002-1244-9126',
                                name = 'Daniel Lins da Silva',
                                identifier = 'daniellins',
                                mbox = 'mailto:daniellins@usp.br',
                                seeAlso = 'http://bioprov.poli.usp.br/students/6613529')

ag_andre <- createAgentProvenance(uri = 'http://orcid.org/0000-0003-4627-0244',
                                name = 'Andre Filipe de Moraes Batista',
                                identifier = 'andrefilipe',
                                mbox = 'mailto:andrefmb@usp.br',
                                seeAlso = 'http://bioprov.poli.usp.br/students/8788438')

experiment <- createExperiment(uri = 'http://bioprov.poli.usp.br/experiment/31', entityType='Experiment',
                                title = 'Species Distribution Modeling of Uroditellus beldingi ',
                                description = 'This work aims to show the relationship between physical aspects and
the occurrences of Uroditellus beldingi in the Western US. We use the R programming language and DISMO Package to
execute this experiment.',
                                wasDerivedFrom = e_projectReport, wasAttributedTo = list(ag_daniel,ag_andre),
                                date = date(), creator = ag_daniel)

allSpeciesData <- read.csv('speciesRecords/Uroditellus_beldingi_training_modern_xy.csv', header=T)

oBeech <- subset(allSpeciesData, SPECIES=='Uroditellus beldingi')

### PROVENANCE - ORIGIN OF OCCURRENCE DATASET (DATA PORTALS, SEARCH PARAMETERS, PARTIAL DATASETS AND DATA CLEANING
ACTIVITIES)

e_arctos <- createEntityProvenance(experiment = experiment,
                                uri = 'http://arctos.database.museum/',
                                identifier = 'ARCTOS',

```

```

        entityType = 'Software',
        title = 'Collaborative Collection Management Solution',
        source = 'http://arctos.database.museum/',
        description = 'Arctos is both a community and a comprehensive collection
management information system. ')

e_manis <- createEntityProvenance(experiment = experiment,
                                uri = 'http://manisnet.org/',
                                identifier = 'MANIS',
                                entityType = 'Software',
                                title = 'Mammal Networked Information System',
                                source = 'http://manisnet.org/',
                                description = 'Information System developed by seventeen North American
institutions with support from NSF.')

e_param_species <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/species',
                                          identifier = "species", entityType = 'InputParameter', type = "Text", value = "Urocitellus beldingi")
e_param_startdate <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/sdate',
                                          identifier = "startDate", entityType = 'InputParameter', type = "Text", value = "1970")
e_param_enddate <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/edate',
                                          identifier = "endDate", entityType = 'InputParameter', type = "Text", value = "2009")

e_occ_ARCTOS <- createEntityProvenance(experiment = experiment,
                                       uri = 'http://bioprov.poli.usp.br/dataset/Urocitellus_beldingi_arctos',
                                       identifier = 'occ_modern_arctos',
                                       entityType = 'Dataset',
                                       scientificName = 'Urocitellus beldingi',
                                       title = 'Occurrence file of the Urocitellus_beldingi species',
                                       numberOfRecords = '111',
                                       source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/speciesRecords/Urocitellus_beldingi_Arctos.csv')

e_occ_MANIS <- createEntityProvenance(experiment = experiment,
                                       uri = 'http://bioprov.poli.usp.br/dataset/Urocitellus_beldingi_manis',

```

```

        identifier = 'occ_modern_manis',
        entityType = 'Dataset',
        scientificName = 'Urocitellus beldingi',
        title = 'Occurrence file of the Urocitellus_beldingi species',
        numberOfRecords = '174',
        source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/speciesRecords/Urocitellus_beldingi_Manis.txt')

act_dataRetrieving1 <- createActivityProvenance(experiment = experiment,
        uri = 'http://bioprov.poli.usp.br/step1_retrieving',
        activityType = 'Acquisition',
        title = 'Retrieving Occurrence Records from ARCTOS Portal',
        decisions = 'Used presence-only records',
        used = list(e_arctos, e_param_species, e_param_startdate,
e_param_enddate),
        generated = e_occ_ARCTOS)

act_dataRetrieving2 <- createActivityProvenance(experiment = experiment,
        uri = 'http://bioprov.poli.usp.br/step2_retrieving',
        activityType = 'Acquisition',
        title = 'Retrieving Occurrence Records from MANIS Portal',
        decisions = 'Used presence-only records',
        used = list(e_manis, e_param_species, e_param_startdate,
e_param_enddate),
        generated = e_occ_MANIS)

e_occ_ready <- createEntityProvenance(experiment = experiment,
        uri = 'http://bioprov.poli.usp.br/dataset/Urocitellus_beldingi_xy',
        identifier = 'occ_modern_training',
        activityType = 'Dataset',
        scientificName = 'Urocitellus beldingi',
        title = 'Occurrence file of the Urocitellus_beldingi species.',
        source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/speciesRecords/Urocitellus_beldingi_training_mode
rn_xy.csv',

```

```

        numberOfRecords = '32',
        wasDerivedFrom = list(e_occ_MANIS, e_occ_ARCTOS))

act_dataMerging <- createActivityProvenance(experiment = experiment,
        uri = 'http://bioprov.poli.usp.br/step3_merging',
        activityType = 'Combination',
        title = 'Merging Occurrence Records from ARCTOS and MANIS',
        used = list(e_occ_ARCTOS, e_occ_MANIS))

act_dataCleaning <- createActivityProvenance(experiment = experiment,
        uri = 'http://bioprov.poli.usp.br/step4_cleaning',
        activityType = 'Preparation',
        title = 'Cleaning Occurrence Records based on criteria described in the
file.',
        description = 'Cleaning Criteria:
https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/ex1_oBeech_modern_randomBg_r/data_cleaning_criteri
a.txt',
        wasInformedBy = act_dataMerging,
        generated = e_occ_ready)

predStackModern <- stack(
    raster('westernUsClimate_modern/BIO02.asc'),
    raster('westernUsClimate_modern/BIO05.asc'),
    raster('westernUsClimate_modern/BIO06.asc'),
    raster('westernUsClimate_modern/BIO07.asc'),
    raster('westernUsClimate_modern/BIO03.asc'),
    raster('westernUsClimate_modern/BIO13.asc'),
    raster('westernUsClimate_modern/BIO14.asc'),
    raster('westernUsClimate_modern/BIO18.asc'),
    raster('westernUsClimate_modern/BIO15.asc')
)

```



```

projection(predStackModern) <- CRS('+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0') # coordinate
reference system for WGS84

plot(predStackModern) # plot predictors

### PROVENANCE - ORIGIN OF ENVIRONMENTAL LAYERS (PREDICTORS AND PRISM DATA PORTAL)

ag_prism <- createAgentProvenance(experiment = experiment,
                                uri = 'http://www.prism.oregonstate.edu/',
                                name = 'PRISM Climate Group',
                                identifier = 'PRISM',
                                homepage = 'http://www.prism.oregonstate.edu/',
                                seeAlso = 'http://www.cefa.dri.edu/Westmap/Westmap_home.php?page=Prism101.php')

e_env_bio02 <- createEntityProvenance(experiment = experiment,
                                     uri = 'http://bioprov.poli.usp.br/map_bioclim_bio02',
                                     identifier = 'bioclim_bio02',
                                     type = 'StillImage',
                                     entityType = 'Map',
                                     title = 'Mean diurnal temperature range',
                                     source =
'https://gitlab.com/daniellins/SDM_Urocyon_vulpinus_beldingi/raw/master/westernUsClimate_modern/BIO02.asc',
                                     description = 'Resolution=800m',
                                     wasAttributedTo = ag_prism)

e_env_bio05 <- createEntityProvenance(experiment = experiment,
                                     identifier = 'bioclim_bio05',
                                     entityType = 'DataResource',
                                     entityFormat = 'RasterFile',
                                     uri = 'http://bioprov.poli.usp.br/map_bioclim_bio05',
                                     title = 'Maximum temperature of the warmest month',
                                     source =
'https://gitlab.com/daniellins/SDM_Urocyon_vulpinus_beldingi/raw/master/westernUsClimate_modern/BIO05.asc',
                                     description = 'Resolution=800m',
                                     wasAttributedTo = ag_prism)

```

```

e_env_bio06 <- createEntityProvenance(experiment = experiment,
                                     uri = 'http://bioprov.poli.usp.br/map_bioclim_bio06',
                                     identifier = 'bioclim_bio06',
                                     entityType = 'DataResource',
                                     entityFormat = 'RasterFile',
                                     title = 'Minimum temperature of the coldest month',
                                     source =
'https://gitlab.com/daniellins/SDM_Urociellus_beldingi/raw/master/westernUsClimate_modern/BIO06.asc',
                                     description = 'Resolution=800m',
                                     wasAttributedTo = ag_prism)

e_env_bio07 <- createEntityProvenance(experiment = experiment,
                                     uri = 'http://bioprov.poli.usp.br/map_bioclim_bio07',
                                     identifier = 'bioclim_bi07',
                                     entityType = 'DataResource',
                                     entityFormat = 'RasterFile',
                                     title = 'Temperature annual range (BIO05-BIO06)',
                                     source =
'https://gitlab.com/daniellins/SDM_Urociellus_beldingi/raw/master/westernUsClimate_modern/BIO07.asc',
                                     description = 'Resolution=800m',
                                     wasAttributedTo = ag_prism)

e_env_bio03 <- createEntityProvenance(experiment = experiment,
                                     uri = 'http://bioprov.poli.usp.br/map_bioclim_bio03',
                                     identifier = 'bioclim_bi03',
                                     entityType = 'DataResource',
                                     entityFormat = 'RasterFile',
                                     title = 'Isothermality (BIO2/BIO07)',
                                     source =
'https://gitlab.com/daniellins/SDM_Urociellus_beldingi/raw/master/westernUsClimate_modern/BIO03.asc',
                                     description = 'Resolution=800m',
                                     wasAttributedTo = ag_prism)

e_env_bio13 <- createEntityProvenance(experiment = experiment,

```

```

        uri = 'http://bioprov.poli.usp.br/map_bioclim_bio13',
        identifier = 'bioclim_bi13',
        entityType = 'DataResource',
        entityFormat = 'RasterFile',
        title = 'Precipitation of Wettest Month',
        source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO13.asc',
        description = 'Resolution=800m',
        wasAttributedTo = ag_prism)

e_env_bio14 <- createEntityProvenance(experiment = experiment,
        uri = 'http://bioprov.poli.usp.br/map_bioclim_bio14',
        identifier = 'bioclim_bi14',
        entityType = 'DataResource',
        entityFormat = 'RasterFile',
        title = 'Precipitation of Driest Month',
        source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO14.asc',
        wasAttributedTo = ag_prism)

e_env_bio18 <- createEntityProvenance(experiment = experiment,
        uri = 'http://bioprov.poli.usp.br/map_bioclim_bio18',
        identifier = 'bioclim_bi18',
        entityType = 'DataResource',
        entityFormat = 'RasterFile',
        title = 'Precipitation of Warmest Quarter',
        source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO18.asc',
        description = 'Resolution=800m',
        wasAttributedTo = ag_prism)

e_env_bio15 <- createEntityProvenance(experiment = experiment,
        uri = 'http://bioprov.poli.usp.br/map_bioclim_bio15',
        identifier = 'bioclim_bi15',
        entityType = 'DataResource',
        entityFormat = 'RasterFile',

```

```

        title = 'Precipitation Seasonality (Coefficient of Variation)',
        source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/westernUsClimate_modern/BIO15.asc',
        description = 'Resolution=800m',
        wasAttributedTo = ag_prism)

dir.create('ex1_oBeech_modern_randomBg_r', recursive=TRUE)

start.time <- Sys.time()
modelBasic <- maxent(
  x=predStackModern,
  p=oBeech[, c('LONG_WGS84', 'LAT_WGS84')],
  path=paste(getwd(), '/ex1_oBeech_modern_randomBg_r', sep=''),
  args=c(
    'randomtestpoints=30',
    'betamultiplier=1',
    'linear=true',
    'quadratic=true',
    'product=true',
    'threshold=true',
    'hinge=true',
    'threads=2',
    'responsecurves=true',
    'jackknife=true',
    'askoverwrite=false'
  )
)
end.time <- Sys.time()
time.taken <- end.time - start.time

# look at model output (HTML page)
modelBasic # note that unlike running Maxent from the program the prediction raster is not automatically created, so
we have to do that in the next step

```

```
### PROVENANCE - DATA ANALYSIS (SOFTWARES, MAXENT PARAMETERS, MODEL)
```

```
e_param_maxent1 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/rtp',
  identifier = "randomtestpoints", entityType = "InputParameter", value = "30")
e_param_maxent2 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/bm',
  identifier = "betamultiplier", entityType = "InputParameter", value = "1")
e_param_maxent3 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/linear',
  identifier = "linear", entityType = "InputParameter", value = "true")
e_param_maxent4 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/quad',
  identifier = "quadratic", entityType = "InputParameter", value = "true")
e_param_maxent5 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/product',
  identifier = "product", entityType = "InputParameter", value = "true")
e_param_maxent6 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/thld',
  identifier = "threshold", entityType = "InputParameter", value = "true")
e_param_maxent7 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/hinge',
  identifier = "hinge", entityType = "InputParameter", value = "true")
e_param_maxent8 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/threads',
  identifier = "threads", entityType = "InputParameter", value = "2")
e_param_maxent9 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/rc',
  identifier = "responsecurves", entityType = "InputParameter", value = "true")
e_param_maxent10 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/jn',
  identifier = "jackknife", entityType = "InputParameter", value = "true")
e_param_maxent11 <- createEntityProvenance(experiment = experiment, uri = 'http://bioprov.poli.usp.br/param/aow',
  identifier = "askoverwrite", entityType = "InputParameter", value = "false")

e_dismo <- createEntityProvenance(experiment = experiment,
  uri = 'https://cran.r-project.org/web/packages/dismo/',
  identifier = 'DISMO',
  entityType = 'software',
  title = 'dismo: Species Distribution Modeling',
  description = 'Functions for species distribution modeling, that is, predicting
entire geographic distributions form occurrences at a number of sites and the environment at these sites.',
  source = 'https://cran.r-project.org/src/contrib/dismo_1.1-1.tar.gz',
  wasAttributedTo = 'http://desp.ucdavis.edu/people/robert-j-hijmans')

e_maxent <- createEntityProvenance(experiment = experiment,
```

```

uri = 'https://www.cs.princeton.edu/~schapire/maxent/',
identifier = 'MAXENT',
entityType = 'Software',
title = 'MAXENT Software',
isVersionOf = '3.3.3',
source = 'https://www.cs.princeton.edu/~schapire/maxent/',
description = 'http://homepages.inf.ed.ac.uk/lzhang10/maxent.html',
wasAttributedTo = 'http://dbpedia.org/resource/Robert_Schapire')

e_rstudio <- createEntityProvenance(experiment = experiment,
uri = 'http://www.rstudio.com/products/rstudio/',
identifier = 'RSTUDIO',
entityType = 'Software',
title = 'RStudio Software',
isVersionOf = '0.99.893',
source = 'http://www.rstudio.com/products/rstudio/download/',
description = 'http://www.rstudio.com/products/rstudio/')

e_model <- createEntityProvenance(experiment = experiment,
uri = 'http://bioprov.poli.usp.br/model/Urocitellus_beldingi',
identifier = 'SDMModel',
entityType = 'DataResource',
entityFormat = 'Archive',
title = 'Resulting Model from MAXENT',
source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/tree/master/ex1_oBeech_modern_randomBg_r/',
wasAttributedTo = ag_daniel, ag_andre)

act_dataAnalysing <- createActivityProvenance(experiment = experiment,
uri = 'http://bioprov.poli.usp.br/step5_analysing',
activityType = 'Analysis',
title = 'Species Distribution Modeling',
decisions = 'From 19 BIOCLIM variables, were kept those are expected to
be biologically meaningful to the species and that had cross-correlations >-0.7 and <0.7 (Tingley and Herman 2009)',
#startedAtTime = as.character.Date(start.time),
#endedAtTime = as.character.Date(end.time),

```

```

        hasDuration = as.character.Date(time.taken),
        computationalArchitectureUsed = 'notebook core i5, 8Gb RAM, SSD 256Gb',
        used = list(e_dismo, e_maxent, e_rstudio, e_occ_ready,
                    e_env_bio02, e_env_bio03, e_env_bio05, e_env_bio06,
e_env_bio07, e_env_bio13, e_env_bio14, e_env_bio15, e_env_bio18,
                    e_param_maxent1, e_param_maxent2, e_param_maxent3,
e_param_maxent4, e_param_maxent5, e_param_maxent6, e_param_maxent7,
                    e_param_maxent8, e_param_maxent9, e_param_maxent10,
e_param_maxent11),

        generated = e_model)

start.time <- Sys.time()
mapModernBasic <- predict(
  object=modelBasic,
  x=predStackModern,
  filename=paste(getwd(), '/ex1_oBeech_modern_randomBg_r/Urocitellus_beldingi_map_modern', sep=''),
  na.rm=TRUE,
  format='GTiff',
  overwrite=TRUE,
  progress='text'
)
end.time <- Sys.time()
time.taken <- end.time - start.time

# look at map
plot(mapModernBasic, main='Modern')

# add species' records
points(oBeech[ , c('LONG_WGS84', 'LAT_WGS84')])

# overlay a shapefile of western states
library(rgdal)
states <- readOGR('shapefiles', 'westernUs_wgs84')
plot(states, add=TRUE)

```

```

### PROVENANCE - DATA PREDICTION AND VISUALIZATION (SOFTWARES, MAXENT PARAMETERS, MODEL)
e_modelmap <- createEntityProvenance(experiment = experiment,
                                     uri = 'http://bioprov.poli.usp.br/mapmodels/Urocitellus_beldingi',
                                     identifier = 'SDM_Map',
                                     type = 'StillImage',
                                     entityType = 'DataResource',
                                     entityFormat = 'RasterFile',
                                     source =
'https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/ex1_oBeech_modern_randomBg_r/Urocitellus_beldingi
_map_modern.tif',
                                     title = 'Raster Map of Species Distribution Modeling',
                                     wasDerivedFrom = e_model,
                                     wasAttributedTo = list(ag_daniel, ag_andre))

act_dataPrediction <- createActivityProvenance(experiment = experiment,
                                                uri = 'http://bioprov.poli.usp.br/step6_prediction',
                                                activityType = 'Prediction',
                                                description = 'Prediction based on the Distribution Modeling and
modern scenario',
                                                title = 'Prediction based on the Distribuion Modeling and ',
                                                #startedAtTime = as.character.Date(start.time),
                                                #endedAtTime = as.character.Date(end.time),
                                                hasDuration = as.character.Date(time.taken),
                                                computationalArchitectureUsed = 'notebook core i5, 8Gb RAM, SSD
256Gb',
                                                used = e_model,
                                                generated = e_modelmap)

getProvenanceOfEntity(uri = experiment, format = 'png')
getProvenanceOfEntity(uri = e_modelmap, format = 'png')
getProvenanceOfExperiment(uri = experiment, format = 'png')

```

Link para acesso ao arquivo original: https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/blob/master/sdm_script.r

APÊNDICE G – Relatório de Proveniência do Estudo de Caso 2 (formato PNG).

Link para acesso ao arquivo original: https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/tree/master/results

APÊNDICE H – Trecho do Relatório de Proveniência do Estudo de Caso 2 (Formato RDF/Turtle).

```
<http://bioprov.poli.usp.br/mapmodels/Urocitellus_beldingi> a prov:Entity ;
    bioprov:entityFormat "RasterFile" ;
    bioprov:entityType "DataResource" ;
    dcterms:created "2017-02-09 21:06:57" ;
    dcterms:identifier "SDMMap" ;
    dcterms:source
<https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/ex1_oBeech_modern_randomBg_r/Urocitellus_beldingi_map_modern.tif> ;
    dcterms:title "Raster Map of Species Distribution Modeling" ;
    dcterms:type "StillImage" ;
    prov:wasAttributedTo <http://orcid.org/0000-0002-1244-9126>,
        <http://orcid.org/0000-0003-4627-0244> ;
    prov:wasDerivedFrom <http://bioprov.poli.usp.br/model/Urocitellus_beldingi> ;
    prov:wasGeneratedBy bioprov:step6_publication .

<http://bioprov.poli.usp.br/dataset/Urocitellus_beldingi_xy> a prov:Entity ;
    bioprov:activityType "Dataset" ;
    dcterms:created "2017-02-09 21:04:46" ;
    dcterms:identifier "occ_modern_training" ;
    dcterms:source
<https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/raw/master/speciesRecords/Urocitellus_beldingi_training_mode
rn_xy.csv> ;
    dcterms:title "Occurrence file of the Urocitellus_beldingi species." ;
    prov:wasDerivedFrom <http://bioprov.poli.usp.br/dataset/Urocitellus_beldingi_arctos>,
        <http://bioprov.poli.usp.br/dataset/Urocitellus_beldingi_manis> ;
    prov:wasGeneratedBy bioprov:step4_cleaning .
```

Link para acesso ao arquivo original: https://gitlab.com/daniellins/SDM_Urocitellus_beldingi/tree/master/results