UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Uma infraestrutura semântica para integração de dados científicos sobre biodiversidade

Kleberson Junio do Amaral Serique

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Kleberson Junio do Amaral Serique

Uma infraestrutura semântica para integração de dados científicos sobre biodiversidade

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Dilvan de Abreu Moreira

USP – São Carlos
Fevereiro de 2018

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S485i Serique, Kleberson Junio do Amaral
 Uma infraestrutura semântica para integração de
dados científicos sobre biodiversidade / Kleberson
Junio do Amaral Serique; orientador Dilvan Abreu
Moreira. -- São Carlos, 2017.
 158 p.

 Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2017.

 1. Web Semântica. 2. Informática para
Biodiversidade. 3. Web de Dados. 4. Linked Data. I.
Abreu Moreira, Dilvan, orient. II. Título.

Kleberson Junio do Amaral Serique

A semantic infrastructure for integrating biodiversity scientific data

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Dilvan de Abreu Moreira

USP – São Carlos
February 2018

*Dedico esse trabalho a minha família, a minha noiva Yasmin
aos meus amigos, ao meu orientador, Dilvan,
aos mestres que me ajudaram a concluir
este trabalho, Laurindo e Andrea, e aos
amigos, Cíntia e Lucas.*

AGRADECIMENTOS

Os agradecimentos principais são direcionados a todos mestres que me conduziram a produzir e crescer na vida acadêmica, em especial Dilvan, meu orientador, Laurindo e Andrea, meus parceiros de pesquisas, aos colegas do grupo de pesquisa Silvio, Flor, João Paulo e Guilherme.

Agradecimentos especiais são direcionados aos meus pais, Dirce e Carlos, minha irmã, Karla, minha noiva, Yasmin, e a todos meus familiares que me deram apoio e compreenderam minhas ausências.

*“Um computador pode ser chamado de inteligente
se ele pode enganar uma pessoa
a pensar que é um ser humano.”
(Alan Turing)*

RESUMO

ABREVIACÃO. Uma infraestrutura semântica para integração de dados científicos sobre biodiversidade. 2018. 158 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Pesquisas na área de biodiversidade são, em geral, transdisciplinares por natureza. Essas pesquisas tentam responder problemas complexos que necessitam de conhecimento transdisciplinar e requerem a cooperação entre pesquisadores de diversas disciplinas. No entanto, é raro que duas ou mais disciplinas distintas tenham observações, dados e métodos em formatos que permitam a colaboração imediata sobre hipóteses complexas e transdisciplinares. Hoje, a velocidade com que qualquer disciplina obtém avanços científicos depende de quão bem seus pesquisadores colaboram entre si e com tecnologistas das áreas de bancos de dados, gerenciamento de workflow, visualização e tecnologias, como computação em nuvem. Dentro desse cenário, a Web Semântica surge, não só como uma nova geração de ferramentas para a representação de informações, mais também para a automação, integração, interoperabilidade e reutilização de recursos. Neste trabalho, uma infraestrutura semântica é proposta para a integração de dados científicos sobre biodiversidade. Sua arquitetura é baseada na aplicação das tecnologias da Web Semântica para se desenvolver uma infraestrutura eficiente, robusta e escalável aplicada ao domínio da Biodiversidade. O componente central desse ambiente é a linguagem BioDSL, uma Linguagem de Domínio Específico (DSL) para mapear dados tabulares para o modelo RDF, seguindo os princípios de Linked Open Data. Esse ambiente integrado também conta com uma interface Web, editores e outras facilidades para conversão/integração de conjuntos de dados sobre biodiversidade. Para o desenvolvimento desse ambiente, houve a participação de instituições de pesquisa parceiras que atuam na área de biodiversidade da Amazônia. A ajuda do Laboratório de Interoperabilidade Semântica do Instituto Nacional de Pesquisas da Amazônia (INPA) foi fundamental para a especificação e testes do ambiente. Foram pesquisados vários casos de uso com pesquisadores do INPA e realizados testes com o protótipo do sistema. Nesses testes, ele foi capaz de converter arquivos de dados reais sobre biodiversidade para RDF e interligar automaticamente entidades presentes nesses dados a entidades presentes na web (nuvem LOD). Num experimento envolvendo 1173 registros de espécies ameaçadas, o ambiente conseguiu recuperar automaticamente 967 (82,4%) entidades (URIs) da LOD referentes a essas espécies, com matching completo para o nome das espécies, 149 (12,7%) com matching parcial (apenas um dos nomes da espécie), 36 (3,1%) não tiveram correspondências (sem resultados nas buscas) e 21 (1,7%) sem registro das espécies na LOD.

Palavras-chave: Web Semântica, Informática para Biodiversidade, Web de Dados, Linked Data.

ABSTRACT

ABREVIACÃO. **A semantic infrastructure for integrating biodiversity scientific data.** 2018. 158 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Research in the area of biodiversity is, in general, transdisciplinary in nature. This type of research attempts to answer complex problems that require transdisciplinary knowledge and require the cooperation between researchers of diverse disciplines. However, it is rare for two or more distinct disciplines to have observations, data, and methods in formats that allow immediate collaboration on complex and transdisciplinary hypotheses. Today, the speed which any discipline gets scientific advances depends on how well its researchers collaborate with each other and with technologists from the areas of databases, workflow management, visualization, and internet technologies. Within this scenario, the Semantic Web arises not only as a new generation of tools for information representation, but also for automation, integration, interoperability and resource reuse. In this work, a semantic infrastructure is proposed for the integration of scientific data on biodiversity. This architecture is based on the application of Semantic Web technologies to develop an efficient, robust and scalable infrastructure for use in the field of Biodiversity. The core component of this infrastructure is the BioDSL language, a Specific Domain Language (DSL) to map tabular data to the RDF model, following the principles of Linked Open Data. This integrated environment also has a Web interface, editors and other facilities for converting/integrating biodiversity datasets. For the development of this environment, we had the participation of partner research institutions that work with Amazon biodiversity. The help of the Laboratory of Semantic Interoperability of the National Institute of Amazonian Research (INPA) was fundamental for the specification and tests of this infrastructure. Several use cases were investigated with INPA researchers and tests were carried out with the system prototype. In these tests, the prototype was able to convert actual biodiversity data files to RDF and automatically interconnect entities present in these data to entities present on the web (LOD cloud). In an experiment involving 1173 records of endangered species, the environment was able to automatically retrieve 967 (82.4%) LOD entities (URIs) for these species, with complete matching for the species name, 149 (12.7%) with partial matching (only one of the species names), 36 (3.1%) with no matching and 21 (1.7%) no have records at LOD.

Keywords: Semantic Web, Informatics for Biodiversity, Data Web, Linked Data.

LISTA DE ILUSTRAÇÕES

Figura 1 – Roteiro para leitura dos Capítulos da Tese	37
Figura 2 – Os três níveis de estudo da diversidade biológica na área de Biologia	39
Figura 3 – O funcionamento do <i>Metacat</i> com seus componentes	48
Figura 4 – Funcionamento da rede do KNB com o <i>Morpho</i> e <i>Metacat</i> integrados	49
Figura 5 – Arquitetura de funcionamento do DiGIR	51
Figura 6 – Arquitetura de funcionamento do protocolo BioCASE	53
Figura 7 – Diagrama da Arquitetura de funcionamento do SinBiota	56
Figura 8 – Arquitetura da rede do <i>speciesLink</i>	57
Figura 9 – Componentes do Programa LBA e suas interligações	62
Figura 10 – O fluxo de dados no LBA-DIS	63
Figura 11 – Focos de pesquisa do Programa LBA	64
Figura 12 – Como o TEAM trabalha	70
Figura 13 – Estações do TEAM do sítio de Manaus	72
Figura 14 – Fontes de dados sobre biodiversidade na Web	73
Figura 15 – Infraestrutura atual para disseminação de dados sobre biodiversidade dos programas e projetos científicos do INPA e MPEG	74
Figura 16 – Arquitetura em camadas da Web Semântica	79
Figura 17 – Exemplo de URI	80
Figura 18 – Representação de uma tripla em RDF	81
Figura 19 – Um grafo RDF após a adição de recursos usando FOAF	81
Figura 20 – Perfis atuais da Linguagem OWL	84
Figura 21 – O diagrama da nuvem de dados presentes na LOD	90
Figura 22 – Visão geral da OntoBio	96
Figura 23 – Configurando o arquivo CSV e mapeando-o para RDF em BioDSL	117
Figura 24 – Executando a BioDSL pelo Terminal	118
Figura 25 – Estrutura de funcionamento da infraestrutura semântica para dados sobre Biodiversidade	127
Figura 26 – Repositório de Arquivos da Infraestrutura	129
Figura 27 – Visão geral do BioDSL Web Editor	130
Figura 28 – Web editor de códigos da infraestrutura	131
Figura 29 – <i>Upload</i> de arquivos de Ontologias	131
Figura 30 – Gerenciador de Prefixos	132
Figura 31 – Abas de URIs com e sem variáveis de ontologia e prefixos	132

Figura 32 – Aba de opções do editor de código BioDSL e de opções da Tabela	133
Figura 33 – Visualizador de arquivos CSV do BioDSL Web Editor	133
Figura 34 – Interface de consulta SPARQL da Infraestrutura	134

LISTA DE CÓDIGOS-FONTE

Código-fonte 1 – BioDSL: Configurando o arquivo CSV	107
Código-fonte 2 – BioDSL: carregando Ontologias e definindo prefixos	108
Código-fonte 3 – BioDSL: Definindo uma URI base	109
Código-fonte 4 – BioDSL: Configurando o triplestore	110
Código-fonte 5 – BioDSL: Função URI	110
Código-fonte 6 – BioDSL: Função Matching	111
Código-fonte 7 – BioDSL: Função Map	113
Código-fonte 8 – Mapeamento de espécies brasileiras ameaçadas	136
Código-fonte 9 – Mapeamento de espécies brasileiras ameaçadas	137
Código-fonte 10 – Mapeamento de dados sobre as eleições paulistas de 2014	139

LISTA DE TABELAS

Tabela 1 – Comparação dos recursos entre as ferramentas de mapeamento CSV para RDF	94
Tabela 2 – Comparação dos recursos da BioDSL com outras ferramentas de mapeamento CSV para RDF	104
Tabela 3 – Palavras Reservadas da BioDSL	115
Tabela 4 – Objetos Reservados da BioDSL	115

LISTA DE ABREVIATURAS E SIGLAS

3G	Gráfico Global Gigante
ABCD	Access Biological Collections Data
BioCASE	The Biological Collection Access Service for Europe
BPS	BioCASE Provider Software
BRAHMS	Botanical Research And Herbarium Management System
CABS	Center for Applied Biodiversity Science
CDB	Convenção sobre Diversidade Biológica
CI	Conservation International
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CNUMAD	Conferência das Nações Unidas sobre Meio Ambiente e Desenvolvimento
CRIA	Centro de Referência em Informação Ambiental
CSS	Cascading Style Sheets
CSV	Comma Separated Values
DiGIR	Distributed Generic Information Retrieval
DSL	Domain-Specific Language
DSL	Domain-Specific Language
DwC	Darwin Core
DwC-A	Darwin Core Archive
EML	Ecological Metadata Language
ER	Entidade de Relacionamento
Fapesp	Fundação de Amparo à Pesquisa do Estado de São Paulo
GBIF	Global Biodiversity Information Facility
GBRDS	Global Biodiversity Resources Discovery System
GEF	Global Environment Facility
GML	Geography Markup Language
GPL	GNU General Public License
GPL	General-Purpose Languages
GWT	Google Web Toolkit
HTML	Hypertext Markup Language
IB	Informática da Biodiversidade ou Informática para Biodiversidade
ICMBIO	Instituto Chico Mendes de Conservação da Biodiversidade

ILTER	International Long Term Ecological Research
INPA	Instituto Nacional de Pesquisas da Amazônia
INPE	Instituto Nacional de Pesquisas Espaciais
IPT	Integrated Publishing Toolkit
JSON	JavaScript Object Notation
JVM	Java Virtual Machine
KNB	Knowledge Network for Biocomplexity
LBA	Programa de Grande Escala da Biosfera-Atmosfera na Amazônia
LBA-DIS	Sistema de Dados e Informações do Programa LBA
LD	Linked Data
LIS	Laboratório de Interoperabilidade Semântica
LME	LBA Metadata Editor
LOD	Linked Open Data
LTER	Long Term Ecological Research Network
MCTIC	Ministério da Ciência, Tecnologia, Inovações e Comunicações
MMA	Ministério do Meio Ambiente
MPEG	Museu Paraense Emílio Goeldi
NASA	National Aeronautics and Space Administration
NBGI	Núcleo de BioGeoInformática
NCEAS	National Center for Ecological Analysis and Synthesis
NEx	Núcleos Executores
NRs	Núcleos Regionais
NSF	National Science Foundation
NSF	National Science Foundation
OGC	Open Geospatial Consortium
OIL	Ontology Inference Layer
ONG	Organização não-governamental
OntoBio	Ontologia de Biodiversidade
ONU	Organização das Nações Unidas
P21C	Plants for the 21st Century
PDBFF	Projeto Dinâmica Biológica de Fragmentos Florestais
PELD	Programa de Pesquisas Ecológica de Longa Duração
PNB	Política Nacional de Biodiversidade
PPBio	Programa de Pesquisas em Biodiversidade
RAP	Rapid Assessment Program
RDF	Resource Description Framework
REST	Representational state transfer

RFD–S	Resource Description Framework – Schema
RIA	Rich Internet Application
SGBD	Sistemas de Gerenciamento de Banco de Dados
SHOE	Simple HTML Ontological Extensions
SI	Smithsonian Institution
SiBBR	Sistema de Informação sobre a Biodiversidade Brasileira
SINBIO	Banco de dados de Inventários Biológicos da Amazônia
SLOSS	Single Large Or Several Small reserves of equal area
SML	Sparqlification Mapping Language
SPARQL	Simple Protocol and RDF Query Language
SQL	Structured Query Language
SUFRAMA	Superintendência da Zona Franca de Manaus
SweoIG	Semantic Web Education and Outreach Interest Group
TAPIR	TDWG Access Protocol for Information Retrieval
TDWG	Biodiversity Information Standards
TEAM	Tropical Ecology, Assessement and Monitoring
TI	Tecnologias e Informação
TSE	Tribunal Superior Eleitoral
UDDI	Universal Description Discovery and Integration
UEFS	Universidade Estadual de Feira de Santana
UEMA	Universidade Estadual do Maranhão
UESC	Universidade Estadual de Santa Cruz
UFAC	Universidade Federal do Acre
UFAM	Universidade Federam do Amazonas
UFMT	Universidade Federal de Mato Grosso
UFOPA	Universidade Federal do Oeste do Pará
UFPA	Universidade Federal do Pará
UFRN	Universidade Federal do Rio Grande do Norte
UFRR	Universidade Federal de Roraima
UFS	Universidade Federal de Sergipe
UFT	Universidade Federal do Tocantins
UML	Unified Modeling Language
UNEMAT	Universidade do Estado de Mato Grosso
UNEP	UN Environment Programe
Unicamp	Universidade Estadual de Campinas
UNIR	Universidade Federal de Rondônia
URI	Uniform Resource Identifier

UUID	Universally Unique Identifie
W3C	World Wide Web Consortium
W3C	World Wide Web Consortium
WD	Web de Dados
WKT	Well-Known Text
WWF	World Wide Fund for Nature
WWW	World Wide Web
WWW	World Wide Web
XML	eXtensivel Markup Language

SUMÁRIO

1	INTRODUÇÃO	29
1.1	Contextualização e Motivação	29
1.2	Justificativa	33
1.3	Hipótese e objetivos	34
1.4	Contribuições ao Estado da Arte	35
1.5	Publicações geradas	36
1.6	Organização da Tese	36
2	INFORMÁTICA PARA BIODIVERSIDADE	39
2.1	Considerações Iniciais	39
2.2	Padrões de dados e metadados	41
2.2.1	<i>Darwin Core – DwC</i>	42
2.2.2	<i>Access Biological Collections Data – ABCD</i>	43
2.2.3	<i>Ecological Metadata Language – EML</i>	44
2.3	Ferramentas de Gestão de Dados sobre Biodiversidade	45
2.3.1	<i>Botanical Research and Herbarium Management System – BRAHMS</i>	45
2.3.2	<i>Specify</i>	46
2.3.3	<i>Metacat</i>	47
2.3.4	<i>Morpho</i>	47
2.4	Protocolos de Disseminação de Dados sobre Biodiversidade	49
2.4.1	<i>Distributed Generic Information Retrieval– DiGIR</i>	49
2.4.2	<i>Biological Collection Access Service – BioCASE</i>	51
2.4.3	<i>TDWG Access Protocol for Information Retrieval – TAPIR</i>	52
2.4.4	<i>Integrated Publishing Toolkit – IPT</i>	53
2.5	Repositórios de dados sobre Biodiversidade	55
2.5.1	<i>Sistema de Informação Ambiental do Programa Biota/Fapesp – SinBiota</i>	55
2.5.2	<i>speciesLink</i>	55
2.5.3	<i>Global Biodiversity Information Facility – GBIF</i>	58
2.5.4	<i>Sistema de Informação sobre a Biodiversidade Brasileira – SiBBR</i>	58
2.6	Programas e Projetos de Pesquisa sobre Biodiversidade na Amazônia	59

2.6.1	<i>Programa de Grande Escala da Biosfera-Atmosfera na Amazônia – LBA</i>	60
2.6.2	<i>Programa de Pesquisas em Biodiversidade – PPBio</i>	65
2.6.3	<i>Programa de Pesquisas Ecológicas de Longa Duração – PELD</i>	68
2.6.4	<i>Projeto Dinâmica Biológica de Fragmentos Florestais – PDBFF</i> . .	69
2.6.5	<i>Projeto de Ecologia, Avaliação e Monitoramento de Florestas Tropicais – TEAM</i>	70
2.7	<i>Análise do Fluxo de Dados sobre Biodiversidade das Pesquisas Amazônicas</i>	71
2.8	<i>Considerações Finais</i>	75
3	WEB SEMÂNTICA: TECNOLOGIAS DE INTEGRAÇÃO DE DADOS	77
3.1	<i>Considerações Iniciais</i>	77
3.2	<i>Arquitetura da Web Semântica</i>	78
3.2.1	<i>Uniform Resource Identifier – URI</i>	79
3.2.2	<i>Resource Description Framework – RDF</i>	80
3.2.3	<i>Ontologias e a Web Ontology Language – OWL</i>	82
3.2.4	<i>SPARQL Protocol and RDF Query Language – SPARQL</i>	85
3.2.5	<i>GeoSPARQL</i>	86
3.2.6	<i>Triplestores</i>	88
3.3	<i>Linked Data – LD</i>	89
3.4	<i>Mapeando arquivos tabulados para o modelo RDF</i>	91
3.5	<i>A Ontologia sobre Biodiversidade</i>	95
3.6	<i>Considerações Finais</i>	97
4	BIODSL: UMA DSL PARA MAPEAMENTO DE DADOS SOBRE BIODIVERSIDADE	99
4.1	<i>Considerações Iniciais</i>	99
4.2	<i>Linguagem de Domínio Específico</i>	101
4.2.1	<i>Linguagem Groovy</i>	102
4.3	<i>BioDSL: uma DSL para mapear dados sobre biodiversidade para o modelo RDF</i>	103
4.4	<i>Comparação com outras ferramentas de mapeamento</i>	104
4.5	<i>Funcionamento da BioDSL</i>	106
4.5.1	<i>Configurando o arquivo CSV na BioDSL</i>	107
4.5.2	<i>Carregando Ontologias e definindo Prefixos</i>	108
4.5.3	<i>Definindo uma URI base</i>	109
4.5.4	<i>Configurando o Triplestore</i>	110
4.5.5	<i>A Função URI</i>	110

4.5.6	<i>A Função Matching</i>	111
4.5.7	<i>A Função Map</i>	113
4.6	Palavras e objetos reservados	114
4.7	Um exemplo de uso da BioDSL	115
4.8	Executando a BioDSL no Terminal	117
4.9	Considerações Finais	118
5	INFRAESTRUTURA SEMÂNTICA PARA DADOS SOBRE BIODIVERSIDADE	121
5.1	Considerações Iniciais	121
5.2	Trabalhos do Grupo de Pesquisa	123
5.3	A extensão da OntoBio	123
5.4	Busca Semântica	124
5.5	O SWI <i>Gazetteer</i>	125
5.6	O funcionamento da Infraestrutura proposta	125
5.7	Experimentos com a Infraestrutura	134
5.7.1	<i>Integrando a lista de espécies ameaçadas brasileiras</i>	134
5.7.1.1	<i>Conversão dos dados do ICMBIO e integração com a DBpedia</i>	135
5.7.1.2	<i>Conversão dos dados do GBIF Backbone Taxonomy</i>	137
5.7.1.3	<i>Interligação dos conjuntos de dados</i>	137
5.7.2	<i>Integrando dados das eleições de 2014</i>	138
5.8	Considerações Finais	139
6	CONCLUSÃO E TRABALHOS FUTUROS	141
6.1	Resumo do trabalho realizado	141
6.2	Parcerias	143
6.3	Contribuições ao Estado da Arte	144
6.4	Publicações geradas até o momento da defesa	145
6.5	Trabalhos futuros	146
	REFERÊNCIAS	147

INTRODUÇÃO

1.1 Contextualização e Motivação

Atualmente, grande parte dos avanços em pesquisas científicas deve-se ao fato de serem resultados de trabalhos transdisciplinares, ou seja, da interação de diversas áreas do conhecimento humano (PENNINGTON, 2007).

Pesquisas transdisciplinares aplicam-se a esforços centrados em problemas do mundo real, elas são orientadas a problemas e integradas ao conhecimento prático (MEHRING *et al.*, 2011). Pesquisas transdisciplinares geram um dilema desorientador, identificado como uma assimetria nos entendimentos de um determinado tópico de pesquisa de uma disciplina em relação a outra (PENNINGTON *et al.*, 2013). Quando esse dilema for mitigado, ou até mesmo equacionado, levará a uma aprendizagem transformadora, conduzindo a uma reestruturação e integração de conceitos, dados e métodos que são fundamentais para a geração da ciência transformadora. Em geral, as pesquisas transformadoras são impulsionadas por idéias que têm o potencial de mudar radicalmente a nossa compreensão de um importante conceito científico ou da engenharia, além de poder resultar na criação de um novo paradigma ou até de um novo campo da ciência ou da engenharia (RANDALL *et al.*, 2007).

Exemplo disso é a área de biodiversidade, considerada como uma área transdisciplinar de pesquisa (MEHRING *et al.*, 2011). As atividades realizadas pelos biólogos compreendem a coleta de dados em campo, a análise de espécimes coletados, a análise dos habitats dos espécimes e a análise das relações entre os espécimes, entre outras. Os dados sobre biodiversidade refletem fatos biológicos relacionados aos demais fatos do conhecimento humano, demonstrando a heterogeneidade explícita e implícita desses dados.

Pesquisas recentes, na área de biodiversidade, vêm sendo conduzidas buscando resolver questões complexas que necessitam de conhecimento transdisciplinar. Além do mais, o desenvolvimento de pesquisas transdisciplinares sobre biodiversidade será fundamental, nas

próximas décadas, garantindo o conhecimento necessário para subsidiar as decisões políticas para conservação (CANHOS, 2003). Normalmente, as pesquisas na área de biodiversidade procuram estudar a variação das espécies e entre espécies ao longo do tempo, além das suas relações com fatores geográficos, ecológicos, temporais e antropogênicos. Ou seja, os estudos sobre biodiversidade são transdisciplinares por natureza e requerem a cooperação entre pesquisadores de diversas disciplinas distintas com o uso intensivo de conhecimentos tácitos e explícitos (ALBUQUERQUE, 2011).

No entanto, é raro que duas ou mais disciplinas distintas tenham observações, dados e métodos que permitam a colaboração imediata sobre hipóteses complexas de caráter transdisciplinar (BENDA *et al.*, 2002).

Atualmente, a velocidade com que qualquer disciplina obtém avanços científicos dependerá de quão bem seus pesquisadores colaborem entre si e com os especialistas da área de Tecnologias da Informação (TI), que lidam com tecnologias como: bancos de dados; gerenciamento de *workflow* científico; visualização da informação; tecnologias de computação em nuvem entre outras.

Nesse contexto, com os recentes avanços tecnológicos, surge uma nova área de estudos denominada de Informática para Biodiversidade ou Informática da Biodiversidade (IB), a qual se refere ao uso de TI para apoiar os estudos sobre biodiversidade, por meio da organização do conhecimento sobre os organismos biológicos e os sistemas ecológicos que os cercam (HOBERN *et al.*, 2013). O objetivo central da IB é desenvolver sistemas que permitam a interoperabilidade e a síntese de conhecimento em grandes matrizes de sistemas locais e incorporá-los em arquiteturas de conhecimento globais, como o *Global Biodiversity Information Facility* (GBIF) (BISBY, 2000). Dessa maneira, o desafio central para a comunidade de IB é fornecer os meios para compartilhar e sintetizar rapidamente os dados e o conhecimento que eles fornecem para construir um mapa global da biodiversidade facilmente acessível e alinhado (GURALNICK; HILL, 2009; BISBY, 2000; CANHOS *et al.*, 2004).

Na área de IB, além do foco no desenvolvimento e evolução de software para a análise e síntese de dados sobre biodiversidade, nota-se, também, avanços significativos na definição de padrões, como o Darwin Core (WIECZOREK *et al.*, 2012), o *Access Biological Collections Data* (Access to Biological Collections Data task group, 2005) e a *Ecological Metadata Language* (FEGRAUS *et al.*, 2005)) e protocolos, como o *Distributed Generic Information Retrieval* (DIGIR) (DIGIR, 2013), *TDWG Access Protocol for Information Retrieval* (TAPIR) (TDWG, 2010) e *Integrated Publishing Toolkit* (IPT) (ROBERTSON, 2011; GBIF, 2013). Eles são usados para integração dos dados sobre biodiversidade distribuídos mundialmente e são fundamentais para a construção da infraestrutura global de informação sobre biodiversidade (CANHOS, 2003).

Uma revolução ocorreu na internet com o surgimento da *World Wide Web* (WWW) (BERNERS-LEE, 1992; BERNERS-LEE *et al.*, 1993), ou simplesmente Web, onde os documentos Web passam a ser o centro desse novo cenário. Documentos Web também fazem ligações

uns aos outros, não importando em que computador eles estejam localizados. Os usuários da Web são capazes de navegar por estes documentos, procurando por informações relevantes. A Web Semântica surge, neste contexto, como uma área proeminente em tecnologias ligadas a integração de dados, ao raciocínio lógico e por permitir a interpretação semântica, por máquinas e seres humanos, de grandes volumes de dados distribuídos pela Web, oriundos de diversas disciplinas. Ela forma uma rede de informações processáveis e derivadas de dados por meio de uma teoria semântica. Essa teoria semântica fornece uma descrição de “significado”, na qual a conexão lógica de termos estabelece a interoperabilidade entre sistemas (SHADBOLT; BERNERS-LEE; HALL, 2006). Neste cenário, computadores são necessários para processar e compreender os volumosos conteúdos dos documentos Web, juntar e interligar informações relevantes, melhorar a precisão das pesquisas na Web, etc.

Segundo Tim Berners-Lee, a Web Semântica apresenta-se como uma nova Web, uma evolução natural da WWW, que visa dar significado aos recursos da Web, permitindo que estes sejam compreendidos por pessoas e por sistemas computacionais (BERNERS-LEE; HENDLER; LASSILA, 2001). Ela também permite a interconexão de conjuntos de dados distintos, por meio do compartilhamento de conceitos ontológicos¹ representados como identificadores universais, como o *Uniform Resource Identifier* (URI). Essa interconexão entre datasets é chamada de *Linked Data* (LD) ou dados vinculados. A interligação dos conteúdos de recursos na Web aos seus conceitos ontológicos torna possível a estruturação do conhecimento por meio de ontologias, formando, assim, grandes grafos de conhecimento acessíveis livremente pela Web, tanto por humanos quanto por máquinas. O benefício é que essas interligações explícitas permitem a análise dos dados vinculados resultantes em uma abordagem transdisciplinar (KAUPPINEN; ESPINDOLA; JONES, 2012).

As ontologias, neste contexto, representam um conjunto de conceitos dentro de um domínio e o relacionamentos destes. Uma ontologia é uma especificação explícita de uma conceitualização². O termo é emprestado da filosofia, onde uma ontologia é um relato sistemático da existência (GRUBER, 1993). Para entender essa definição sobre ontologia, deve ser claro o que é uma Conceitualização. Conceitualização é uma interpretação estruturada de uma parte do Mundo que as pessoas usam para se comunicar e pensar sobre o Mundo (BORST, 1997). Para um biólogo, tal conceitualização pode definir, por exemplo, que os animais classificam-se em grupos chamados de espécies e que os animais pertencentes a uma espécie têm hábitos alimentares semelhantes. Com base nesses hábitos alimentares, as espécies podem ser sub-categorizadas em herbívoros, carnívoros e onívoros.

As ontologias também podem ser definidas como uma especificação formal e explícita de uma conceitualização compartilhada (BORST, 1997). Nessa definição, uma especificação formal significa ser legível por computadores; uma especificação explícita diz respeito a conceitos,

¹ são aqueles que dizem respeito ao “ser”, a “identidade”, a “consistência” e a “existência” das diferentes entidades

² é uma visão abstrata e simplificada do mundo que se deseja representar

atributos, relações, restrições e axiomas que são explicitamente definidos; compartilhado quer dizer conhecimento consensual; e conceitualização diz respeito a um modelo abstrato de algum fenômeno do mundo real (ALBUQUERQUE, 2016).

A Web Semântica, as ontologias e o LD, juntos, configuram uma plataforma promissora para publicação de dados de estudos científicos, podendo incluir dados sobre biodiversidade. Essas tecnologias dispõem de uma estrutura simples para representar e integrar dados, o *Resource Description Framework* (RDF) (W3C, 2004b).

O RDF é um modelo simples de dados composto por afirmações na forma de triplas que representam um recurso, uma propriedade e um objeto, que tornam o RDF uma linguagem altamente escalável. O uso do URI para identificar entidades como autores, instituições e artigos científicos é um instrumento valioso para a identificação de descrições repetidas dessas entidades (tanto redundantes como complementares), principalmente quando elas são provenientes de fontes distintas de dados.

A evolução das tecnologias da Web Semântica, das Ontologias e de LD permitem que dados na Web sejam representados e interligados em um modelo compreensível por máquinas, surgindo, assim, a Web de Dados (WD), conhecida também como *Giant Global Graph* (3G). Dessa forma, um novo paradigma é criado, onde os dados na Web tendem a tornar-se uma Web de Dados (teia de dados), ou seja, um ambiente de acesso livre com diversas informações interligadas entre si, baseadas, principalmente, em conteúdos de dados abertos de diversas disciplinas, como a bioinformática, química, medicina, etc.

Entretanto, dados sobre biodiversidade, disponíveis na Web, estão em formatos que não permitem uma rápida colaboração com outras áreas de conhecimento, tanto para descobertas de novos conhecimentos, a partir da análise dos relacionamentos entre as entidades (e compreensão transdisciplinar dos fatos que as descrevem), como, também, para serem reutilizados e relacionados com outras áreas do conhecimento.

Além do mais, grande parte dos dados sobre biodiversidade na Web estão disponíveis em formatos tabulados de dados, como *Comma Separated Value* (CSV), associados à padrões de metadados. No entanto, embora pareça apenas uma simples conversão de formato de dados (CSV para o RDF), este processo não é trivial, pois deve envolver o uso compartilhado de conceitos ontológicos, além do compartilhamento de URIs sobre entidades biológicas usadas por todos os conjuntos de dados, ou seja, as entidades biológicas deverão ser vistas como um recurso URI, na Web, por todos os registros de dados que as referenciam, criando, dessa forma, interoperabilidade entre os sistemas que utilizam estes dados.

Uma Infraestrutura Semântica de Dados sobre Biodiversidade poderá explorar ao máximo as tecnologias da Web Semântica, melhorando assim não só a disseminação, mais, também, a qualidade dos dados sobre biodiversidade, por meio da análise semântica dos fatos em relação ao conhecimento formalizado nas ontologias, além de poder apontar possíveis erros semânticos

presentes nos dados, como locais de ocorrência sobre biodiversidade fora do *habitat* de uma determinada espécie e até garantir a desambiguação de registros de nomes de espécies, que não são mais utilizados atualmente (nomenclaturas desatualizadas de registros de dados antigos, ou em discussão na área de biodiversidade).

Além do mais, embora os dados de outras disciplinas estejam participando efetivamente da WD, ainda há poucos dados sobre biodiversidade. A criação de um ambiente computacional que integre dados sobre biodiversidade com dados da WD permitirá que estes dados possam ser utilizados em pesquisas transdisciplinares futuras, facilitando uma rápida compreensão dos fatos presentes nos dados, que transcendem suas disciplinas, por meio da análise e consultas em um gráfico de conhecimento global, como o 3G ou WD.

Essa problemática caracteriza a motivação para o desenvolvimento desta pesquisa. Esta tese apresenta uma Infraestrutura Semântica para Dados sobre Biodiversidade, tendo as tecnologias da Web Semântica, como plataforma para interoperabilidade, e os princípios de LD, como mecanismo de interligação dos dados sobre biodiversidade com a WD.

1.2 Justificativa

Diversas discussões apontam para uma integração em larga escala de dados científicos (PENNINGTON, 2007; COUNCIL *et al.*, 2010; OVASKA *et al.*, 2010). Nas últimas décadas, tanto os financiadores de pesquisas científica quanto os próprios pesquisadores reconheceram que dados são uma parte muito importante da produção da pesquisa e que, portanto, merecem uma melhor atenção com sua disponibilidade e governança.

O uso de recursos computacionais é imprescindível para se fazer ciência no atual cenário mundial dos dados, das informações e dos conhecimentos científicos. Dessa maneira, surge a IB com o objetivo de desenvolver sistemas que permitam a interoperabilidade, a síntese de conhecimento e a difusão de dados em arquiteturas de conhecimento globais. Por outro lado, as tecnologias da Web Semântica vêm proporcionando o surgimento da WD, que é um 3G baseado em um modelo simples de representação de conhecimento altamente escalável, o RDF.

No entanto, apesar de grandes esforços da área de IB no desenvolvimento e manutenção de padrões de dados e metadados sobre biodiversidade e na publicação e disseminação, a nível global, desses dados, ainda não é satisfatório o grau de interoperabilidade entre os mesmos com as demais áreas do conhecimento humano, demandando esforços individuais a cada adição de dados de outras disciplinas, além de uma notória discreta participação desse tipo de dado na WD (WIECZOREK *et al.*, 2012).

Esse problema gera limitações no desenvolvimento de estudos transdisciplinares sobre dados de biodiversidade. Ele torna a procura por respostas a questões de pesquisas complexas uma atividade nada trivial que necessita de diferentes conjuntos de dados multidisciplinares. Além

disso, grande parte dos dados sobre biodiversidade possui um elevado grau de conhecimento transdisciplinar, que muitas vezes não são entendidos de forma explícita.

Logo, permitir a integração dos dados sobre biodiversidade com dados de outras disciplinas correlacionadas, por meio da semântica dos dados e metadados contido nos mesmos, pode proporcionar uma colaboração quase que imediata, explicitando o conhecimento transdisciplinar e facilitando: o reuso dos dados; a adoção de padrões e a diminuição dos esforços repetitivos e custosos na realização de novas coletas de dados; etc.

Para se chegar em uma integração colaborativa de dados científicos de diversas áreas do conhecimento humano, a Web Semântica se apresenta como uma solução proeminente. Muitas disciplinas, como as ciências da vida e saúde, vêm utilizando essas tecnologias para proporcionar integração semântica entre os seus dados (ANTEZANA; KUIPER; MIRONOV, 2009). No entanto, ainda é discreta a participação dos dados sobre Biodiversidade.

Prover dados sobre Biodiversidade na WD, seguindo os princípios de LD e de forma colaborativa, constitui a principal motivação e justificativa para a presente pesquisa.

1.3 Hipótese e objetivos

Esta pesquisa tem como foco a integração semântica de dados sobre biodiversidade utilizando as tecnologias e padrões da Web Semântica definidos pelo *World Wide Web Consortium* (W3C), que é a principal organização de padronização da WWW. A hipótese desta pesquisa é:

“A interoperabilidade semântica irá propiciar que dados científicos sejam consultados e consumidos de forma transdisciplinar, no campo da biodiversidade ou em qualquer outro campo do conhecimento humano.”

O objetivo geral deste trabalho é o desenvolvimento de uma infraestrutura computacional para integração semântica de dados científicos sobre biodiversidade. Para tanto, foram utilizados os padrões e tecnologias da Web Semântica para criar um ambiente colaborativo e transdisciplinar onde usuários podem interligar seus conjuntos de dados a conjuntos de dados de outros usuários e com a WD, para explorarem questões científicas que extrapolem o domínio dos seus dados.

O desenvolvimento deste trabalho foi focado nos dados sobre biodiversidade da Amazônia, fornecidos por grupos parceiros do Laboratório de Interoperabilidade Semântica (LIS) do Instituto Nacional de Pesquisas da Amazônia (INPA), como também os dados disponíveis em infraestruturas globais de disseminação de dados sobre biodiversidade, como o GBIF e *speciesLink*.

Esta infraestrutura proposta é baseada nos conceitos de LD, no uso de ontologias sobre biodiversidade e uso de padrões e tecnologias da Web Semântica. Foram alcançados os seguintes objetivos específicos:

- Identificar as estruturas de dados, padrões de dados, modelos de dados e metadados e ontologias que são utilizados pelos projetos científicos amazônicos sobre biodiversidade (parceiros do LIS) para, então, associar a eles padrões e tecnologias da Web Semântica;
- Desenvolver ou re-usar tecnologias que permitam criar novos grafos de conhecimento com os conjuntos de dados sobre biodiversidade (mapear esses dados para o modelo RDF);
- Permitir a reutilização de identificadores para entidades ligadas à biodiversidade (URI), nos novos grafos de conhecimentos (URIs já associadas a essas entidades em outros conjuntos de dados na WD).
- Identificar/estender características formais de ontologias para a integração semântica dos dados científicos, quando necessário;
- Desenvolver um protótipo da infraestrutura semântica para integração de dados científicos de biodiversidade que diminua o grau de complexidade da integração semântica para os usuários finais (biólogos);
- Desenvolver e integrar ferramentas e APIs necessárias para a integração de dados com base na infraestrutura proposta;
- Realizar a avaliação do protótipo da infraestrutura semântica com usuários finais: os pesquisadores de biodiversidade do LIS/INPA.

1.4 Contribuições ao Estado da Arte

Algumas contribuições científicas desta pesquisa são elencadas a seguir:

- Desenvolvimento de mecanismos automatizados para a reutilização de URIs de entidades conhecidas nos conjuntos de dados e na WD, por meio de buscas semânticas em *endpoints* SPARQL (*entity matching on linked open data*);
- Desambiguação de nomes de localidades presentes nos dados de biodiversidade utilizando-se um *Gazetteer* semântico;
- Uma nova abordagem baseada no uso de Linguagens de Domínio Específico, a Linguagem BioDSL, para mapear dados tabulados para o modelo RDF, seguindo os princípios de LD, uso de Ontologias, *reasoning* e publicação dos dados em *triplestores*;
- Desenvolvimento de protótipo funcional para edição colaborativa via Web de *scripts* BioDSL para produção de bases de conhecimentos e disseminação dos dados via tecnologias da Web Semântica;

- Desenvolvimento e integração dos componentes citados acima em uma infraestrutura semântica proposta, como: servidor de aplicações, *triplestores*, editores Web para arquivos tabulados e *scripts* BioDSL.

1.5 Publicações geradas

Foram geradas as seguintes publicações durante o desenvolvimento desse doutorado:

- O artigo “*BioDSL: a domain-specific language for mapping and dissemination of biodiversity data in the LOD*” foi apresentado no *X Brazilian e-Science Workshop* do XXXVI Congresso da Sociedade Brasileira de Computação (SERIQUE *et al.*, 2016).
- O artigo “*SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data. Future Generation Computer Systems*” foi publicado no periódico *Future Generation Computer Systems*, fator de impacto 3,997 (CARDOSO *et al.*, 2015).
- O artigo “*Improving Biodiversity Data Retrieval through Semantic Search and Ontologies*” foi apresentado no *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (AMANQUI *et al.*, 2014).
- O artigo “*A Gazetteer for Biodiversity Data as a Linked Open Data Solution*” foi apresentado no *2014 IEEE 23rd International WETICE Conference* (CARDOSO *et al.*, 2014).
- O capítulo de livro “*Mo Porã: Uma ferramenta para o gerenciamento distribuído de repositórios de dados científicos na web*” foi publicado no livro *Cenários* (SERIQUE; SANTOS; ALBUQUERQUE, 2014).
- O artigo “*Semantic Search Architecture for Retrieving Information in Biodiversity Repositories*” foi apresentado no *6th Seminar on Ontology Research in Brazil* (AMANQUI *et al.*, 2013).

1.6 Organização da Tese

A hierarquia dos capítulos desta tese é apresentada na Figura 1. Nessa figura, setas indicam o fluxo dos assuntos para orientação da leitura. Este documento está organizado em seis Capítulos.

O Capítulo 2 apresenta um referencial teórico sobre a área de IB, destacando os padrões de dados e metadados, as ferramentas de gestão de dados, os repositórios de dados. Suplementarmente, são apresentados os programas e projetos científicos apoiados pelos nossos parceiros, o LIS/INPA. A análise do funcionamento destes projetos mostrou o fluxo dos dados sobre biodiversidade que vão da geração, gestão, armazenamento até sua disseminação.

Fonte: Elaborada pelo autor.

Figura 1 – Roteiro para leitura dos Capítulos da Tese

No Capítulo 3 é apresentado o referencial teórico sobre a Web Semântica, destacando as tecnologias e padrões de dados, como: URI, RDF, OWL, SPARQL, etc. Essas tecnologias são a base para o desenvolvimento do objetivo geral deste trabalho.

O Capítulo 4 apresenta a linguagem BioDSL, uma *Domain-Specific Language* (DSL) para o mapeamento de dados sobre biodiversidade para o modelo RDF com integração com a WD.

O Capítulo 5 apresenta a infraestrutura semântica proposta por meio da implementação de um protótipo. Ele também apresenta as ligações entre os referenciais teóricos dos Capítulos 2 e 3, demonstrando como os conceitos foram integrados no desenvolvimento do protótipo da Infraestrutura Semântica Computacional.

O Capítulo 6 apresenta o resumo do trabalho realizado, as parcerias firmadas durante a tese, as contribuições da tese ao estado da arte, as publicações geradas pela tese e os trabalhos futuros.

Por fim, são apresentadas as referências bibliográficas utilizadas nesta tese.

INFORMÁTICA PARA BIODIVERSIDADE

2.1 Considerações Iniciais

O termo biodiversidade se refere ao estudo global da vida na Terra, levando em consideração a sua variação intrínseca. Nesse contexto, a diversidade biológica pode ser considerada em três níveis: (1) molecular; (2) organismo e (3) ecológico ([SARKAR, 2007](#)). A Figura 2 mostra que os estudos em bioinformática se concentram principalmente no desenvolvimento de hipóteses ao nível molecular, já os estudos em IB enfatizam principalmente a organização do conhecimento nos níveis mais altos (organismos e ambientes).

A situação da biodiversidade, neste início de século, pode ser caracterizada pela combinação de processos acelerados de destruição de ecossistemas primários, associados a esforços mobilizadores para a conservação e uso sustentável da biodiversidade, além de grandes avanços em TI ([SANTOS *et al.*, 2011](#)). Essa combinação de fatores está propiciando o surgimento de uma nova área de desenvolvimento científico e tecnológico, denominada IB (ou no inglês *Biodiversity Informatics*) ([CANHOS, 2003](#)).

A IB é uma disciplina relativamente nova que se estende da ciência da computação, no

Fonte: Adaptada de [Sarkar \(2007\)](#).

Figura 2 – Os três níveis de estudo da diversidade biológica na área de Biologia

contexto dos dados sobre biodiversidade (ARIÑO; CHAVAN; KING, 2011).

O desafio central, enfrentado pela comunidade de IB, está em fornecer meios para compartilhar e sintetizar rapidamente dados e conhecimentos para a construção de uma visão global da biodiversidade. Para tanto, é imprescindível o desenvolvimento de sistemas de informação que proporcionem interoperabilidade de dados e síntese do conhecimento, com base em grandes matrizes de sistemas locais para incorporar estes dados à arquiteturas de conhecimento globais, como o GBIF (BISBY, 2000; CANHOS *et al.*, 2004; GURALNICK; HILL, 2009).

Nesse sentido, a IB representa a conjunção do gerenciamento e uso eficiente de informações sobre biodiversidade com o desenvolvimento de novas ferramentas computacionais para análise e compreensão dessas informações. Para Hobern, a IB refere-se ao uso da TI para apoiar essas necessidades, por meio da organização de conhecimento sobre organismos biológicos e os sistemas ecológicos que os formam (HOBERN *et al.*, 2013).

Em agosto de 2017, redes de informação sobre biodiversidade, como o GBIF, BioCase e *speciesLink*, juntas forneciam cerca de 791 milhões de registros de ocorrência de biodiversidade, dos quais cerca de 130 milhões desses registros pertencem a espécimes físicos mantidos em coleções de museus e universidades (GÜNTSCH *et al.*, 2017).

A comunidade de IB tem mantido também a atenção em relação a descrição e acesso a informações complementares, que podem estar associadas a uma amostra específica de um espécime ou registro da ocorrência do mesmo. A importância destas informações complementares está na ampliação do escopo dos dados potencialmente relevantes, que incluem uma ampla gama de medidas observadas sobre os aspectos bióticos e abióticos do ambiente. Por exemplo, ao analisar padrões na abundância global de certos táxons, outras informações, fora do escopo da biodiversidade, podem ser parâmetros importantes à serem analisados, como informações sobre as geadas, os tipos de solos, o uso da terra, etc.

Atualmente, dados sobre a biodiversidade relacionados à diferentes campos do conhecimento estão disponíveis na Web para todos os interessados (cidadãos, políticos, governos, iniciativa privada, etc.) e com uma forte demanda para integrar, sintetizar e visualizar esta informação para diferentes fins e por diferentes tipos de usuários finais. Dessa maneira, a necessidade de integração da comunidade de IB converge com a de outros domínios do conhecimento, como das ciências ambientais, que dependem de dados multidisciplinares para a compreensão integrada ou holística dos dados (ALBUQUERQUE, 2011).

Os repositórios de dados sobre biodiversidade, em geral, contém dados e informações, tais como medições de instrumentos, registros de espécimes biológicos e observações, que podem conter julgamentos de especialistas sobre os conceitos e classificações de espécies (COSTELLO *et al.*, 2013). Além do mais, os dados de biodiversidade, assim como os dados de muitas outras disciplinas, muitas vezes não são gerados automaticamente por máquinas ou sensores. Dados sobre espécies, por exemplo, são baseados em observações e interpretações humanas e requerem

curadoria constante para se manterem atualizados (COSTELLO *et al.*, 2013).

A integração de diversas fontes de informações digitais é, de fato, um dos grande desafios para a IB. Essa área é constantemente confrontada por inúmeros provedores de dados, cada um com suas próprias comunidades de usuários específicos, assim como o tipo de informação que lhes interessa é bastante diversificada, incluindo nomes, conceitos taxonômicos, espécimes em coleções biológicas, registros de ocorrência de biodiversidade, publicações científicas, genômica e dados de imagens fenotípicas (PAGE, 2008).

A Seção 2.2 apresenta os principais tecnologias da IB no que se referem aos padrões de dados e metadados, ferramentas de gestão, protocolos de disseminação e repositórios de dados na Web. A seção 2.6 apresenta projetos e programas de pesquisas suportados pelos parceiros desta pesquisa, LIS/INPA, que permitiram identificar os padrões de dados e metadados, ferramentas e repositórios utilizados nas pesquisas amazônicas sobre a biodiversidade, bem como uma análise do fluxo dos dados (produzidos e disseminados) por meio das tecnologias da IB adotada por estes projetos.

2.2 Padrões de dados e metadados

O principal desafio dos repositórios de dados sobre biodiversidade é lidar com o caráter altamente interdisciplinar e heterogêneo, tanto dos dados como dos metadados, padronizados ou não padronizados, que podem conter informações que abrangem desde genes até ecossistemas (BACH *et al.*, 2012).

Embora muitos cientistas reconheçam a necessidade de trocar informações eletronicamente, apenas recentemente houve um crescimento na conscientização sobre a necessidade do emprego de padrões de dados e metadados para intercambio de informações sobre biodiversidade (MORRIS, 2013).

Além do mais, os dados sobre biodiversidade podem se distinguir entre si por diversos fatores, como: uso de diferentes protocolos experimentais; modelagem dos dados; protocolo de coleta (que deve ser adequado às particularidades de cada investigação); etc. Esta variação resulta no aumento da heterogeneidade dos dados, que torna-os difíceis de compartilhar e de identificar as co-relações existentes entre eles.

Por sua vez, os metadados, ou dados sobre os dados, podem descrever os atributos do conteúdo de documentos, atividades e até um conjunto de dado inteiro. Em ecologia, por exemplo, os metadados são as informações que descrevem: quem; o que; onde; quando; o porquê; e como um conjunto de dados foi coletado (FEGRAUS *et al.*, 2005). Os metadados podem, também, levar os usuários a ter uma compreensão completa dos conjunto de dados e das características dos seus atributos (SANTOS; Campos dos Santos, 2003).

Atualmente, existem diversas soluções para amenizar os problemas relacionados a inte-

gração de dados sobre biodiversidade, o mais importante é o uso de vocabulários controlados para representarem dados e metadados. Nas seções 2.2.1, 2.2.2, 2.2.3 são abordados as tecnologias relacionada aos padrões de dados e metadados sobre a biodiversidade, como: *Darwin Core*, *Access Biological Collections Data* e *Ecological Metadata Language*.

2.2.1 *Darwin Core – DwC*

O *Darwin Core* (DwC) é um vocabulário, ou um conjunto de termos, para descrever dados sobre biodiversidade (ROBERTSON *et al.*, 2014). O *Darwin Core* foi baseado em um outro padrão, o *Dublin Core*, proposto inicialmente para metadados sobre obras escritas e objetos digitais em geral, como vídeos, sons, imagens, etc. (HEIDORN, 2011). O objetivo principal do DwC foi criar uma linguagem comum para partilhas de dados sobre ocorrências de biodiversidade, de forma complementar, reutilizando padrões de metadados de outros domínios, sempre quando for possível (WIECZOREK *et al.*, 2012). Dessa maneira, o DwC facilita o intercâmbio de informações, entre bancos de dados distribuídos, sobre a ocorrência geográfica de espécies e sobre a existência de espécimes coletados e mantidos em coleções biológicas.

Em 1999 o DwC foi apresentado pela primeira vez, ainda como um conjunto de termos vagamente definidos. O DwC foi progredindo por várias iterações, influenciadas por grupos diferentes, que resultaram em diversas variantes. No entanto, foi necessário uma administração e controle de mudanças centralizados, garantido sua efetividade para integração de dados. Nesse contexto, foi criado o *Darwin Core Task Group*, mantido pelo *Biodiversity Information Standards* (TDWG), que no ano de 2009 lançou o DwC como um padrão oficial.

A filosofia adotada no desenvolvimento do DwC era de manter o padrão o mais simples e aberto, quanto possível, e criar novos termos só quando existir uma grande demanda compartilhada pela comunidade. Além disso, o DwC possui uma história relativamente longa de desenvolvimento baseado em comunidade e é amplamente empregado como um padrão para troca de informações adotado por diversas redes de disseminação de dados sobre biodiversidade, como o *speciesLink* e o GBIF (WIECZOREK *et al.*, 2012).

A documentação do padrão DwC é mantida pelo *Darwin Core Task Group*, que contém vários documentos que estão disponíveis pelas páginas Web. O DwC possui um glossário de termos, que podem ser chamados de propriedades, elementos, campos, nome de colunas, atributos ou conceitos, dependendo do contexto ao qual forem empregados. Estes termos fornecem definições para referência, comentários e exemplos de uso (Darwin Core Task Group, 2009).

O táxon é a figura central do DwC e está relacionado com as informações da localização da sua ocorrência na natureza, documentação das observações humana da mesma, as amostras de espécimes mantida em museus ou em universidades, e outras informações relacionadas.

O DwC têm a capacidade de compartilhar registros com propriedades que não se repetem (de forma hierárquica), como arquivos de texto simples ou no formato *eXtensivel Markup Lan-*

guage (XML). Existe, também, o *Darwin Core Archive* (DwC-A), uma combinação de arquivos tabulados no formato CSV e um documento em XML, que descreve a semântica das colunas do arquivo CSV e as suas relações. O DwC-A é um meio fácil e popular para o compartilhamento de dados sobre biodiversidade. Sua popularização é impulsionada por sua adoção pelas novas ferramentas de publicação de dados, apresentadas na seção 2.4 (WIECZOREK *et al.*, 2012).

Existe, também, uma representação do DwC no formato RDF¹, que inclui a descrição de todos os termos, seus históricos, as relações em si e com termos de outros padrões, em um único arquivo RDF. Embora o DwC seja definido em um documento RDF, a integração dos dados sobre biodiversidade na Web Semântica está em seus estágios iniciais.

Um dos principais desafios para DwC no contexto da Web Semântica é a falta de uma ontologia bem definida, no que diz respeito a uma definição formal de relações entre termos em um domínio definido (WIECZOREK *et al.*, 2012). Uma ontologia definirá as relações entre conceitos como as entidades biológicas, os eventos que documentam onde e quando essas entidades ocorreram e os processos pelos quais essas entidades são identificados como um conceito taxonômico. Sem as relações rigorosas entre os conceitos e as propriedades que os definem, as conexões entre os dados sobre biodiversidade e a informação semanticamente rica relacionada, como a literatura e os genomas, será difícil integrar os conceitos e os dados sobre biodiversidade (WIECZOREK *et al.*, 2012). Isso cria obstáculos à pesquisa semântica interdisciplinar, como na comunidade de LD².

Os repositórios de dados sobre biodiversidade (seção 2.5) e os programas/projetos (seção 2.6), que serão apresentados, adotam amplamente o padrão DwC. Logo, compreender o funcionamento do padrão DwC irá auxiliar no processo de integração dos dados sobre biodiversidade, além de fornecer os subsídios necessários para o mapeamento desses termos com termos das ontologias de biodiversidade, que serão utilizadas em nossa proposta de infraestrutura semântica para os dados de biodiversidade.

2.2.2 *Access Biological Collections Data – ABCD*

Em todas as coleções biológicas existentes no mundo, há um conjunto de atributos em comuns, incluindo elementos específicos das amostras, como o táxon, o número de espécimens e o sexo. Cada item da coleção documenta a existência de uma espécie em um determinado local em um determinado momento (HOLETSCHEK *et al.*, 2012). No entanto, o conjunto de atributos adotados para descrever estes itens variam de coleção para coleção.

O *Access Biological Collections Data* (ABCD) é uma especificação de padrão para os atributos de dados sobre coleções biológicas, que incluem espécimes vivos e preservados, juntamente com as observações de campo que não produziram amostras de comprovação. De forma geral, o ABCD é destinado ao intercâmbio e integração de dados sobre as coletas

¹ <http://rs.tdwg.org/dwc/rdf/dwcterms/history>

² <http://linkeddata.org>

e observações de biodiversidade ([Access to Biological Collections Data task group, 2005](#); [ALBUQUERQUE, 2011](#)).

Para tanto, o ABCD fornece um conjunto de nomes (termos), bem como suas definições, para que pesquisadores e curadores de museus os adotem na definição e no gerenciamento dos dados de suas coleções biológicas. Os objetivos iniciais do desenvolvimento da especificação do padrão ABCD almejavam a abrangência e a generalidade. Dessa maneira, os elementos e conceitos foram criados de modo à fornecer a maior compatibilidade possível com os demais esquemas de dados adotados pelas coleções biológicas (ABCD, 2012). A especificação de dados do ABCD é disponibilizada como um XML *schema* (TDWG, 2009). Este *schema* é uma maneira de expressar dados em uma estrutura portátil, facilitando o seu compartilhamento pela Web.

O DwC e o ABCD são amplamente utilizados nas pesquisas dos programas e projetos apresentados na seção 2.6. Logo, compreender o funcionamento destes padrões irá auxiliar o processo de integração semântica proposto neste trabalho.

2.2.3 Ecological Metadata Language – EML

Ecological Metadata Language (EML) é uma linguagem para formalizar e padronizar o conjunto de conceitos essenciais para a descrição de dados ecológicos ([FEGRAUS *et al.*, 2005](#)). A EML é uma linguagem baseada em XML *schema*, desenvolvida para a comunidade de ecologia para descrever dados ecológicos ([San Gil *et al.*, 2009](#)).

EML surgiu de um código-fonte aberto, produto do esforço de uma comunidade que envolve pesquisadores em ecologia, gestores de informação e desenvolvedores de software, liderados pelo *National Center for Ecological Analysis and Synthesis* (NCEAS) e pela *Long Term Ecological Research Network* (LTER). A EML foi formalmente adotada pela LTER em 2001 ([MAYERNIK; BATCHELLER; BORGMAN, 2011](#)).

Esta linguagem foi destinada ao uso por ecologistas ou por administradores de informações ecológicas. Além de abranger um conjunto de aspectos essenciais sobre dados ecológicos, tais como: os nomes e definições das variáveis; as unidades de medida; a data; a hora; o local de coleta de dados; a identidade da pessoa que coletou os dados; a amostragem; etc.

Atualmente, a EML é desenvolvida como parte da *Knowledge Network for Biocomplexity* (KNB), que corresponde a uma rede internacional de compartilhamento de dados que integra dados de diferentes sítios, laboratórios e pesquisadores ao redor do mundo³.

A forma como a EML é utilizada pelos seus usuários irá determinar o potencial de suas funcionalidades. Por um lado, quando o usuário fornece apenas as informações mínimas exigidas e compatíveis com as regras da EML (como o título, proprietário e ponto de contato), então a funcionalidade desse conjunto de metadados permanece meramente num nível bibliográfico. Por outro lado, quando o usuário fornece informações mais abrangentes e detalhadas, então o

³ <https://knb.ecoinformatics.org>

potencial de funcionalidades é muito mais rico, só então a análise dos dados poderá ser mediada por máquinas. Dessa forma, a EML reduz ambiguidade e incerteza, formalizando os conceitos de metadados em um conjunto abrangente e padronizado de termos e definições destinados especificamente para dados ecológicos.

Assim como o ABCD e o *Darwin Core*, a linguagem EML, também, é amplamente utilizada nos repositórios de dados de biodiversidade dos programas e projetos apresentados na seção 2.6. Logo, compreender o funcionamento da EML irá nos auxiliar no processo de integração semântica e no mapeamento dos seus termos com os termos provenientes de ontologias de biodiversidade, que serão utilizadas para gerar triplas RDF, permitindo, dessa maneira, construir uma semântica necessária para processos automatizados e compreensão dos dados, tanto por humanos como por máquinas, melhorando a qualidade dos dados por localizar inconsistências semânticas, antes não detectadas com as tecnologias atuais de disseminação de dados de biodiversidade.

2.3 Ferramentas de Gestão de Dados sobre Biodiversidade

2.3.1 *Botanical Research and Herbarium Management System – BRAHMS*

O *Botanical Research and Herbarium Management System* (BRAHMS) é um sistema de informação flexível, desenvolvido para integrar e administrar dados sobre pesquisas botânicas, tais como: imagens de espécimes; observações de campo; coleções vivas; bancos de sementes e literatura (BRAHMS, 2013). O BRAHMS é desenvolvido associado ao componente de conservação da iniciativa *Plants for the 21st Century* (P21C) do *Department of Plant Sciences* da Universidade de Oxford. A atividade chave do P21C é desenvolver formas inovadoras para armazenar, analisar e divulgar os grandes volumes de dados de diversidade sobre plantas que são necessários para apoiar a conservação e promover maior ação ambiental.

Este sistema de informação fornece ferramentas computacionais para visualizar, editar, selecionar, consultar, produzir relatórios e mapas, exportar e publicar os dados sobre as pesquisas botânicas, além de contar com os seguintes recursos: gerenciamento de coleções, um módulo de parcelas amostrais, um módulo de banco de sementes, gerenciamento de coleções vivas, gerenciamento de imagens e documentos, cálculo e mapeamento de diversidade e um gerenciador de publicações. Ainda, conta com a possibilidade de trabalhar em rede, permitindo o acesso simultâneo de múltiplos usuários.

Algumas coleções de herbários pertencentes ao Programa de Pesquisas em Biodiversidade (PPBio) utilizam o BRAHMS para gerenciar os dados de seus herbários. Com a infraestrutura semântica proposta neste trabalho, os usuários do BRAHMS poderão interligar seus dados com outros dados sobre biodiversidade, ou não, para ter uma visão mais ampla dos fatos contidos nos seus dados. Para tanto, precisamos estudar e entender como o BRAHMS gerencia

os seus dados e como eles poderão ser extraídos para serem mapeados para nossa infraestrutura semântica.

2.3.2 *Specify*

O *Specify* é uma plataforma de banco de dados de espécies e espécimes para coleções de pesquisas biológicas com recursos como: rastreamento de amostras de espécimes, vínculo de fotografias aos registros dos espécimes e publicação dos dados dos espécimes na internet ([SPECIFY, 2013](#)).

O projeto *Specify* é mantido com fundos dos EUA, do *National Science Foundation* (NSF), e é da *University of Kansas*. O projeto *Specify* conta com diversas ferramentas, tais como:

- *Specify 6* – um software robusto para plataformas Linux, Mac e Windows;
- *Specify 7* – uma versão de ambiente Web do *Specify*;
- *Specify Cloud* – um ambiente de nuvem do *Specify*;
- *Specify Web Portal* – publica os dados das coleções, mapas e imagens para os usuários na Web; e
- *Specify Insight* é um aplicativo móvel para o *Apple iPad* para acessar dados, mapas e imagens em qualquer lugar.

A publicação de dados pela internet é feita por meio do *Specify Web Portal*. Esse portal conta com uma interface visual do *Specify*, denominada *Schema Mapper*, para organizar cópias de saída de registros de espécimes e exportar dados para outros formatos necessários para disseminação para outros projetos. Por exemplo, o *Schema Mapper* organiza registros de espécimes como entradas para o *Integrated Publishing Toolkit* (IPT), que produz arquivos no padrão DwC para disseminação pela rede do GBIF.

Para manter uma identificação permanente dos registros de dados biológicos, o *Specify* implementa identificadores exclusivos globais como o uso de *Universally Unique Identifier* (UUID).

O *Specify* é desenvolvido na tecnologia *Java*⁴, que é portátil para vários sistemas operacionais, como Linux, Mac e Windows ([SPECIFY, 2013](#)). O *Specify* suporta ligações à Web services e extensões para criação de *Java plug-ins*. O *Specify* utiliza o MySQL como sistema gerenciador de banco de dados. Em novembro de 2016 o *Specify* lançou a versão 2.3 do *schema* do seu banco de dados. Tanto o *Specify*, como Java e MySQL possuem licenças abertas de software.

O *Specify* é utilizado para gerenciar algumas coleções do PPBio.

⁴ <http://www.java.com/>

2.3.3 *Metacat*

O *Metacat* é um repositório de dados e metadados que ajuda os cientistas a encontrar, entender e usar efetivamente os conjuntos de dados, seus ou de outros cientistas. O *Metacat* foi desenvolvido pelo *National Center for Ecological Analysis and Synthesis* (NCEAS) da *University of California*.

Os usuários do *Metacat* podem armazenar, consultar e recuperar documentos no formato XML como se fossem *schemas* arbitrários de sistemas de banco de dados relacionais (JONES *et al.*, 2001). O *Metacat* usa metadados estruturados, representados em XML, como o EML, para fornecer descrições ricas de conteúdo de dados em estruturas hierárquicas.

Atualmente, milhares de conjuntos de dados são documentados de forma padronizada e armazenados nos sistemas *Metacat*, fornecendo à comunidade científica uma ampla gama de dados científicos que podem ser facilmente pesquisados, comparados, mesclados ou utilizados de outras maneiras.

A Figura 3 apresenta o funcionamento do *Metacat*.

O *Metacat* possui interfaces de comunicação externa com sistemas de banco de dados (consultar, pesquisa e armazenar dados), sistema de arquivo (armazenamento de documentos e figuras) e serviços de autenticação (como LDAP). Sua interface com usuário é por meio do protocolo HTTP, como nos sistemas Web.

O *Metacat* é utilizado pelo PPBio, Pesquisas Ecológicas de Longa Duração e Projeto Dinâmica Biológica de Fragmentos Florestais para disseminação de dados ecológicos. Boa parte dos usuários da infraestrutura semântica proposta utilizam o *Metacat* para disseminar dados em rede. Uma análise do funcionamento de ferramentas e redes de disseminação de dados é essencial para se capturar requisitos inerentes a estes ambientes.

2.3.4 *Morpho*

O *Morpho* é um aplicativo para o gerenciamento de dados ecológicos, projetado para ajudar pesquisadores na gestão de coleções heterogêneas de dados ecológicos (HIGGINS; BERKLEY; JONES, 2002). O usuário deste sistema pode criar, editar e gerenciar metadados em EML associados a tabelas de dados (planilhas eletrônicas). Ademais, o *Morpho* é útil para cientistas individuais que necessitam gerir seus próprios dados de pesquisa (FEGRAUS *et al.*, 2005).

Além de permitir que os pesquisadores em ecologia descrevam seus dados, o *Morpho*, também, permite o compartilhamento público dos dados ou por meio de colaboradores específicos, usando a rede do KNB por exemplo (por meio do software *Metacat*).

Os principais recursos do *Morpho* incluem (HIGGINS; BERKLEY; JONES, 2002):

Fonte: Adaptada de Jones *et al.* (2001).

Figura 3 – O funcionamento do *Metacat* com seus componentes

- Criação e edição de metadados flexíveis, usando a sintaxe do XML, para viabilizar as trocas de metadados;
- Editor de XML que é configurável usando vários XML e DTD;
- Conformidade com a linguagem EML;
- Interface do tipo *wizard*, para a entrada de metadados;
- Realiza a extração automática de metadados, enquanto importa os dados;
- Máquina de buscas em rede (*Metacat*) ou localmente;
- Controle de versão de dados e metadados;
- Compatível com diversos sistemas operacionais, como *Linux*, *Mac* e *Windows*.

A KNB é uma rede criada para facilitar a pesquisa ecológica e ambiental em bio-complexidade, permitindo a descoberta, o acesso, a interpretação, a integração e a análise de dados ecológicos

Fonte: Adaptada de Jones *et al.* (2001).

Figura 4 – Funcionamento da rede do KNB com o *Morpho* e Metacat integrados

complexos a partir de um conjunto altamente distribuído de estações de campo, laboratórios, centros de pesquisa e pesquisadores individuais (HIGGINS; BERKLEY; JONES, 2002).

A Figura 4 apresenta uma visão geral de funcionamento dos componentes que integram a rede do KNB. O *Morpho* aparece na parte esquerda da Figura 4 e se conecta ao *Metacat* via o protocolo HTTP (Web). A rede KNB conta, também, com uma interface Web onde os usuários podem realizar consultas sobre os dados e metadados compartilhados pela rede.

O *Morpho*, o *Metacat* e o EML, juntos, formam a base da rede KNB e são ferramentas muito utilizadas no gerenciamento de informação em pesquisas ecológicas, que incluem dados sobre biodiversidade.

2.4 Protocolos de Disseminação de Dados sobre Biodiversidade

2.4.1 Distributed Generic Information Retrieval– DiGIR

A primeira rede de coleções biológicas a desenvolver um sistema distribuído foi a rede *Species Analyst*, com o uso do protocolo Z39.50, no final dos anos 90 (HEIDORN, 2011; CANHOS *et al.*, 2005). No entanto, para coleções biológicas, esse protocolo provou ser muito complexo, exigindo muitas adaptações por parte dos provedores de dados (CANHOS *et al.*, 2005).

Para suprir a necessidade de um protocolo mais relacionado aos dados de biodiversidade, iniciou-se uma discussão sobre o desenvolvimento de novo protocolo que fosse mais simples e atendesse a demanda de uma rede distribuída de dados de coleções biológicas. Membros do TDWG, *University of Kansas* e pesquisadores da *University of California* e *California Academy of Sciences* optaram por desenvolver esse protocolo de forma cooperativa e colaborativa, então publicaram os primeiros códigos fonte no *SourceForge* (um ambiente para desenvolvimento de softwares de código aberto). No Brasil, a equipe do Centro de Referência em Informação Ambiental (CRIA), relacionada ao desenvolvimento do *speciesLink*, participou deste desenvolvimento colaborativo. Assim, surgiu o protocolo *Distributed Generic Information Retrieval* (DiGIR). A proposta original deste protocolo incluiu os seguintes requisitos e objetivos (CANHOS *et al.*, 2004):

- Utilizar padrões e protocolos abertos, como o HTTP, XML e *Universal Description Discovery and Integration* (UDDI);
- Possuir uma separação clara entre: protocolo, software e semântica;
- Ser de fácil instalação e configuração por parte dos provedores de dados;
- Ter o desenvolvimento de forma colaborativa, seguindo os conceitos de código livre;
- Distribuir livremente o produtos por meio da licença pública *GNU General Public License* (GPL).

A arquitetura típica de uma rede DiGIR, apresentada na Figura 5, envolve pelo menos três componentes distintos (CANHOS *et al.*, 2005):

- **Camada de apresentação** – corresponde ao software que interage com o usuário, oferecendo uma interface amigável para especificação de buscas e exibição dos resultados. A camada de apresentação comunica-se com a camada seguinte;
- **Camada de distribuição de mensagens (portal)** – corresponde ao software que recebe requisições da camada de apresentação e as distribui para cada um dos provedores de dados pertencentes à rede. A comunicação com os provedores é feita por meio do protocolo DiGIR.
- **Provedor** – corresponde ao software responsável por receber requisições do portal e traduzi-las para a linguagem de consulta utilizada pelos sistemas de banco de dados do provedor. O processo de tradução da busca no DiGIR inclui o mapeamento que o provedor fez em relação a um ou mais esquemas conceituais utilizados pelas redes em que atua.

Quando a consulta é feita, o aplicativo cliente DiGIR envia a consulta para o provedor DiGIR de cada instituição. A consulta é então traduzida para um pedido equivalente e compatível

Fonte: Adaptada de [DiGIR \(2013\)](#).

Figura 5 – Arquitetura de funcionamento do DiGIR

com o banco de dados local. Dessa maneira, uma resposta para a pesquisa pode ser processada, mesmo que outros detalhes do banco de dados sejam suprimidos, isso permite ter uma visão virtual e uniforme dos conteúdos dos nodos da rede. A velocidade da conexão e a disponibilidade é a maior preocupação do DiGIR ([GODDARD et al., 2011](#)).

O desenvolvimento do DiGIR foi financiado pela NSF dos Estados Unidos e pela Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp).

2.4.2 *Biological Collection Access Service – BioCASE*

O *Biological Collection Access Service (BioCASE)* é um protocolo de disseminação de dados sobre biodiversidade que surgiu da necessidade de viabilizar a troca de dados utilizando esquemas conceituais mais complexos, neste caso o padrão ABCD. O BioCASE surgiu da modificação do protocolo DiGIR, de código aberto. No entanto, essas modificações resultaram na incompatibilidade entre os protocolos *BioCASE* e DiGIR.

O protocolo *BioCASE* é responsável por definir as formas de consulta, o empacotamento dos dados e como as respostas das consultas serão entregues.

O *BioCASE Provider Software (BPS)* é um *middleware* de ligação de dados baseado em XML, usado como uma camada de abstração a frente de bancos de dados relacionais.

O BPS é composto por uma coleção de ferramentas para ser instalada e configurada em um servidor Web e funciona como um provedor de dados.

O principal componente do BPS é o *PyWrapper*, um software desenvolvido na linguagem de programação *Python*.

O BPS requer a versão 2.5, ou superior, da linguagem *Python*.

A conexão com os diferentes bancos de dados relacionais é feita por módulos, que usam *templates* SQL para facilitar o desenvolvimento dos módulos a cada novo banco de dados. Com a reutilização destes *templates* garante o acesso padronizado a uma variedade de Sistemas de Gerenciamento de Banco de Dados (SGBD).

O *PyWrapper* foi projetado para lidar com qualquer tipo de esquema de banco de dados, desde que exista um arquivo de configuração para cada SGBD e para o seu respectivo esquema de banco de dados. Os arquivos de configuração são chamados de PSF e CMF, como apresentado na Figura 6. Esses arquivos de configuração são usados para mapear os atributos do esquema do banco de dados aos conceitos definidos para um esquema conceitual em XML (como o ABCD). Depois de configurado, o BPS é acessível como um serviço do protocolo BioCASE (via HTTP) e pode, também, ser utilizado para criar sistemas de informação heterogêneos distribuídos.

O Portal BioCASE ⁵ é um aplicativo Web que realiza buscas e consultas em provedores de dados BioCASE e em arquivos do padrão *Darwin Core*. Este portal suporta dois tipos de pesquisas, a pesquisa simples e a pesquisa avançada. A pesquisa simples permite ao usuário realizar pesquisas com base em nomes científicos das espécies. A pesquisa avançada permite ao usuário realizar pesquisas com base em nomes científicos, nomes comuns, nomes de coletores, ou nomes das famílias das espécies. Este portal permite, também, que os usuários filtrem as suas pesquisa pelas instituições, coleções, países, localidades e conteúdos multimídia.

2.4.3 TDWG Access Protocol for Information Retrieval – TAPIR

O GBIF financiou um estudo, em 2004, com a intenção de desenvolver um novo protocolo, que atendesse às necessidades das redes DiGIR e das redes BioCASE (DÄÖRING; GIOVANNI, 2004). Um novo protocolo, denominado de *TDWG Access Protocol for Information Retrieval* (TAPIR), foi desenvolvido pelo *Taxonomic Databases Working Group* do TDWG. O objetivo do TAPIR era integrar os protocolos DiGIR e BioCASE, buscando facilitar o processo de integração destes diferentes sistemas (PEIXOTO *et al.*, 2006).

O TAPIR estende recursos dos protocolos, DiGIR e BioCASE, para obter um método genérico de realizar o intercâmbio de informações entre as redes DiGIR e BioCASE (GODDARD *et al.*, 2011). O TAPIR é utilizado como um método de coleta de dados do GBIF, que continua a manter e adicionar funcionalidade ao projeto (GODDARD *et al.*, 2011). O TAPIR, também, é uma recomendação de protocolo de disseminação de dados de 2010 do TDWG (TDWG, 2010).

⁵ <http://search.biocase.org/europe/>

Fonte: Adaptada de [BioCASE \(2013\)](#).

Figura 6 – Arquitetura de funcionamento do protocolo BioCASE

Por ser um protocolo genérico, o TAPIR pode ser utilizado por diferentes portais de informação. Isto permite a criação de múltiplos recursos de dados, tornando possível recuperar informações em diferentes esquemas de dados. O protocolo TAPIR possui as vantagens de ter seu código fonte aberto, além de não ter qualquer dependência com os sistemas operacionais (multiplataforma), com os esquemas de dados ou com o modelo de saída de dados.

O *TapirLink* é a implementação do protocolo TAPIR, desenvolvido na linguagem de programação PHP ([CORRÊA et al., 2013](#)). O *TapirLink* fornece um ponto de acesso em conformidade com o protocolo TAPIR. Cada ponto de acesso é um serviço separado, podendo ser utilizado para realizar consultas sobre os recursos subjacentes ou a uma fonte de dados local.

2.4.4 Integrated Publishing Toolkit – IPT

Apesar da evolução dos protocolos DiGIR, BioCASE e TAPIR, alguns problemas permaneciam, como dificuldades técnicas de manter os *middlewares* em funcionamento e demora no processamento de respostas por estes sistemas de mensagens distribuídas. Ainda, estes problemas agravam-se em relação do aumento contínuo de novos nodos provedores de dados, elevando os custos para manter em funcionamento uma infraestrutura desse porte ([ROBERTSON et al., 2014](#)). A solução foi implementar uma ferramenta simples e independente de plataforma, que poderia ser facilmente gerenciada pelo corpo técnico das instituições e alavancar padrões de metadados, como o DwC e EML ([ROBERTSON et al., 2014](#)).

O IPT foi desenvolvido a partir de um replanejamento de como a disseminação de dados, no domínio das coleções biológicas, deveriam ocorrer. O IPT é uma ferramenta de software livre desenvolvida na tecnologia *Java* (multiplataforma), utilizada para publicar e compartilhar, por meio da rede do GBIF, conjuntos de dados sobre biodiversidade provenientes de banco de dados,

planilhas eletrônicas e arquivos de texto usando padrões abertos como o *Darwin Core* e EML (GBIF, 2013).

O IPT trouxe, também, para a comunidade de IB a criação do DwC-A, uma remodelagem de como o DwC deveria ser armazenado e compartilhado pelos publicadores de dados em um formato de arquivo em comum (ROBERTSON *et al.*, 2014). O IPT suporta dois subconjuntos de DwC, o *Occurrence Core* e o *Taxon Core*, outras customizações além do padrão DwC são chamados de “extensões” pelo IPT.

O IPT é uma ferramenta Web desenvolvida para usuários proverem, por meio da internet ou uma Intranet, dados sobre biodiversidade, tais como (ROBERTSON, 2011):

- **Registros de ocorrência de biodiversidade:** residem em bancos de dados ou em arquivos de texto, tais como arquivos *Comma Separated Values* (CSV) ou valores separados por vírgulas. Esses dados são normalmente registros de amostras individuais ou de observações de indivíduos de uma determinada espécie.
- **Dados de *checklist* taxonômicos:** residem em bancos de dados ou em arquivos de texto, e correspondem a toda informação associada, tais como a cobertura geográfica ou nomes vernáculos. Esse tipo de conteúdo pode ser usado para expor um conjunto de dados, onde a unidade central é uma espécie.
- **Conjuntos de dados descritivos de alto nível (metadados):** são os dados de autoria capturados por meio da interface Web do IPT, que servem para descrever o conjunto de dados como um todo, incluindo: a taxonomia, a geografia, o escopo temporal, como o dado foi produzido, os requisitos de citação dos dados, entre outros.

O IPT trabalha em conjunto com o *Global Biodiversity Resources Discovery System* (GBRDS), um registrador de dados implementado como um serviço Web *RESTful*, para facilitar a descoberta e recuperação de dados compartilhados pela rede do GBIF. O IPT comunica-se com o GBRDS para registrar os conjuntos de dados das instituições associadas, dessa maneira, permite a publicação e compartilhamento de dados por intermédio da rede do GBIF. O GBRDS fornece, também, novas “extensões” para o IPT, como vocabulários controlados que podem ser utilizados no mapeamento de dados arbitrários.

A ferramenta IPT destina-se a implantação institucional, em vez da implantação individual por cientistas. O IPT complementa outras ferramentas que foram criadas pelo GBIF para ajudar os usuários de dados a explorar, organizar e mobilizar conjuntos de dados sobre biodiversidade (ROBERTSON, 2011). O IPT foi lançado em 2009 e atualmente é o “*wrapper*” (software que empacota dados e transmite pela Web) recomendado pelo GBIF, porém vários *wrappers* legados, como o DiGIR, o PyWrapper (BioCASE) e o TapirLink (TAPIR), ainda estão em uso pelos provedores de dados da rede do GBIF.

2.5 Repositórios de dados sobre Biodiversidade

2.5.1 *Sistema de Informação Ambiental do Programa Biota/Fapesp – SinBiota*

O Sistema de Informação Ambiental do Programa Biota/Fapesp, ou SinBiota, é um sistema de informação brasileiro que congrega dados de coletas e listas de espécies associadas à base cartográfica do estado de São Paulo (CANHOS, 2003). O objetivo do SinBiota é integrar informações geradas pelos pesquisadores vinculados ao Programa Biota/Fapesp (BIOTA, 2013) e relacioná-las a uma base cartográfica digital de qualidade, provendo assim, mecanismos de difusão de informação sobre a biodiversidade paulista para a comunidade científica, os tomadores de decisão, os formuladores de políticas ambientais e os educadores (SINBIOTA, 2013).

O SinBiota foi concebido em 1999 e desenvolvido através de uma parceria entre o CRIA, o Instituto Florestal, os Institutos de Computação e de Geociências e a Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas (Unicamp), responsáveis pela digitalização da base cartográfica do estado de São Paulo.

O acesso ao sistema é permitido somente à pesquisadores cadastrados e associados a algum projeto do programa. Os usuários do SinBiota podem inserir, corrigir ou até remover seus dados do sistema central. A entrada de dados é realizada por meio de um formulário padrão de coleta com campos de dados obrigatórios e opcionais, usando vocabulários controlados mantidos pelos coordenadores de cada projeto. Para segurança e controle do sistema, o acesso é protegido por senhas mantidas pelos coordenadores de cada projeto. Foi, também, desenvolvido uma estrutura de banco de dados que pudesse integrar os dados de todos os grupos taxonômicos.

A Figura 7 apresenta o diagrama da arquitetura de funcionamento do SinBiota. O usuário pode interagir no SinBiota por meio da sua interface Web. O sistema conta com um banco de dados relacional e um servidor de mapas.

O SinBiota é um exemplo de uma iniciativa nacional de sucesso, devido ao grande número de participantes e de informações mantidas por ele. No entanto, ele abrange somente dados do estado de São Paulo e não está vinculado diretamente a nenhum programa e projeto científico que selecionamos. Abordamos o SinBiota, nesta Seção, por ser uma referência nacional no que diz respeito a sistemas de informação voltados a disseminação e integração de dados sobre biodiversidade.

2.5.2 *speciesLink*

O *speciesLink* é um sistema distribuído que integra, em tempo real, dados primários de coleções científicas (CANHOS *et al.*, 2004; SPECIESLINK, 2013). O objetivo inicialmente do *speciesLink* era integrar dados dos acervos de coleções científicas do estado de São Paulo.

Fonte: Adaptada de [Canhos et al. \(2005\)](#).

Figura 7 – Diagrama da Arquitetura de funcionamento do SinBiota

A filosofia do projeto era integrar dados das coleções biológicas, interferindo o mínimo possível na rotina, adaptando os sistemas de informação que já haviam sido previamente adotados. Para tanto, foram listados os sistemas de informação adotados para o gerenciamento das coleções biológicas, como: SinBiota, BRAHMS, *Specify*, Microsoft Access, planilhas eletrônicas e sistemas gerenciadores de banco de dados.

Para realizar as trocas de informações, a equipe do *speciesLink* colaborou no desenvolvimento do protocolo DiGIR, discutido na seção 2.4.1.

Nem todas as coleções biológicas possuem acesso a internet, ou não contam com recursos técnicos capazes de manter um sistema de informação *on-line* na Web, então o *speciesLink* criou servidores regionais, que espelham (replicam) os dados das coleções que possuem estas dificuldades. Para viabilizar esta arquitetura do *speciesLink*, sua equipe desenvolveu interfaces e programas capazes de ler os registros das coleções e atualizar os bancos de dados nos servidores regionais.

A Figura 8 apresenta a arquitetura de funcionamento da rede *speciesLink*. Existem dois níveis de conectividade:

- **O nível com conectividade rápida e estável** – o portal do *speciesLink* conecta-se diretamente aos *Provider* da rede instalados nos servidores regionais e em instituições com recursos Web próprios;

Fonte: Adaptada de Canhos *et al.* (2005).

Figura 8 – Arquitetura da rede do *speciesLink*

- **O nível de conectividade lenta ou instável** – compostos pelos servidores regionais, que por sua vez, recebem conexões das coleções por meio de um agente, o *spLinker*, que extrai e envia as informações das coleções para a disseminação pela rede *speciesLink*.

O *speciesLink* permite a criação de filtros, oferecendo aos curadores das coleções a liberdade de omitir dados (considerados sensíveis), ou seja, o controle total sobre o que será, ou não, disponibilizado pela rede *speciesLink*.

Em 2017, a rede *speciesLink*⁶ contava com 460 coleções e sub-coleções, que juntas oferecem cerca de 8,885,418 registros *on-line*, sendo que 4,047,030 desses registros são georreferenciados (dados com registros de localização geográfica).

Algumas coleções biológicas do PPBio, seção 2.6.2, disseminam dados pela rede *speciesLink*. No entanto, existem muitos registros de outras coleções que estão relacionadas a Amazônia dentro da rede *speciesLink*. Esses registros podem ser úteis para a interligação com os dados das coleções do INPA. Para a integração de dados proposta neste trabalho, os usuários poderão interligar esses dados da rede *speciesLink* com os seus dados dentro da infraestrutura proposta e assim poder explorar um volume maior de informações relacionadas aos seus dados.

⁶ <http://splink.cria.org.br>

2.5.3 Global Biodiversity Information Facility – GBIF

A criação do GBIF partiu da recomendação de 1999 do *Biodiversity Informatics Subgroup* do *Megascience Forum*, criado pela *Organization for Economic Co-operation and Development* (OECD)⁷ (CANHOS *et al.*, 2004). A conclusão dos debates foi “Um mecanismo internacional é necessário para tornar os dados e a informação da biodiversidade acessíveis em todo o mundo”. Em 2001, a recomendação foi assegurado pelo OECD e o GBIF foi oficialmente estabelecido pelos governos dos países participantes.

Atualmente, o GBIF é uma infraestrutura internacional de dados abertos sobre a biodiversidade mantida por governos de diversos países (CANHOS *et al.*, 2004; YESSON *et al.*, 2007). O GBIF opera por meio de uma rede de nodos pela internet, alimentando um portal e serviços Web centralizados para busca de informações sobre a biodiversidade mundial (EDWARDS, 2000).

Em Fevereiro de 2004, o portal do GBIF ficou *on-line* para acesso a dados de coleções, herbários e dados de observações (EDWARDS, 2004).

Os dados disponíveis por meio do GBIF estão relacionados a evidências sobre mais de 1,6 milhão de espécies, coletadas ao longo de três séculos, que incluem observações atuais de cientistas cidadãos, pesquisadores e programas automatizados de monitoramento. Os dados presentes no GBIF podem ser utilizados, dentre outras coisas, para revisões taxonômicas, modelagem de nicho ambiental, compilação de listas de espécies ameaçadas e avaliação de biodiversidade (YESSON *et al.*, 2007).

O GBIF é parceiro de diversas outras iniciativas de disseminação de dados sobre biodiversidade, como *Catalogue of Life*, TDWG, *Consortium for the Barcode of Life* (CBOL), *Encyclopedia of Life* (EOL) e *Global Earth Observation System of Systems* (GEOSS).

Muitos países participantes do GBIF vem criando portais nacionais para melhor informar seus cidadãos e fomentar políticas sobre sua própria biodiversidade, com o uso das ferramentas, códigos abertos e dos dados disponíveis gratuitamente por meio da rede do GBIF. O Brasil, neste cenário, vem desenvolvendo o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr), apresentado na seção 2.5.4.

2.5.4 Sistema de Informação sobre a Biodiversidade Brasileira – SiBBr

Atualmente, existe um grande interesse para a criação de um banco de dados nacional sobre biodiversidade. Esta é uma iniciativa do Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC), denominado de SiBBr (PYLRO *et al.*, 2013).

Este sistema de informação tem o objetivo de integrar informações sobre a biodiversidade e os ecossistemas brasileiros, atualmente dispersos em bancos de dados de diversos órgãos governamentais, universidade, museus e outras fontes espalhados pelo Brasil e pelo Mundo

⁷ <http://www.oecd.org/>

(GADELHA *et al.*, 2014). O intuito é criar um sistema *on-line* com informações de qualidade, tanto para servir ao desenvolvimento das pesquisas científicas, quanto para fundamentar políticas públicas (SIBBR, 2013).

O SiBBR é um sistema *on-line* para consulta livre de informações. Pretende integrar informações sobre a biodiversidade e os ecossistemas brasileiros proveniente de diversas fontes, nacionais e estrangeiras, para subsidiar pesquisas e apoiar os tomadores de decisões na criação e implementação das políticas públicas relacionadas com a biodiversidade do país.

O sistema considera todos os dados disponíveis em instituições como herbários e coleções biológicas, zoológicas e microbianas, no entanto, a iniciativa encontra-se nos primeiros passos e muito trabalho ainda está por ser feito (SETTE; PAGNOCCA; RODRIGUES, 2013).

Este sistema, será o nó brasileiro da rede do GBIF. O SiBBR utiliza os mesmos padrões para colher e agregar dados sobre coleções biológicas, ocorrências e listas de espécies de instituições do Brasil ou do exterior, que publiquem dados sobre a biodiversidade brasileira (GADELHA *et al.*, 2014). Além de integrar dados socioeconômicos, disponibiliza ferramentas de computação de alto desempenho para modelagem de distribuição de espécies. (GADELHA *et al.*, 2014).

O SiBBR é apoiado tecnicamente pelo *UN Environment Programme* (UNEP) e financeiramente pelo *Global Environment Facility* (GEF) (SETTE; PAGNOCCA; RODRIGUES, 2013).

O SiBBR é organizado de forma modular, onde boa parte da funcionalidade está disponível por meio de serviços Web. Os dois principais componentes para gerenciamento de dados são providos pelo Portal de Espécies e Ocorrências e pelo Portal de Dados Ecológicos. Cada instituição pode instalar e cadastrar uma instância do IPT, o portal de Espécies e Ocorrências, por meio do HIT, irá realizar a coleta e indexação dos dados sobre ocorrências de biodiversidade.

O SiBBR é um passo no futuro no que diz respeito aos dados sobre biodiversidade brasileira, uma vez que esta pesquisa visa a construção de infraestrutura semântica, os resultados poderão ser úteis no desenvolvimento do SiBBR.

2.6 Programas e Projetos de Pesquisa sobre Biodiversidade na Amazônia

O Brasil é um país mega-biodiverso, pois abriga em seu território uma das maiores diversidades biológicas do mundo. Por conseguinte, o governo brasileiro vem tomando diversas medidas em relação a proteção e exploração sustentável da biodiversidade brasileira em conformidade com a Convenção sobre Diversidade Biológica (CDB) (CBD, 2013; Secretariat of the Convention on Biological Diversity, 2005). Na Amazônia, diversos projetos e programas científicos vem sendo financiados e mantidos pelo governo brasileiro. Essas iniciativas vem

contribuindo para o melhor entendimento da Amazônia e para a melhoria do processo de tomada de decisão sobre políticas públicas, como leis e criações de novas reservas biológicas, além de contribuir na formação de recursos humanos qualificados em cursos de pós-graduação.

Boa parte das pesquisas transdisciplinares na Amazônia são executadas por duas instituições proeminentes do MCTIC, o INPA e Museu Paraense Emílio Goeldi (MPEG). Essas instituições executam programas e projetos científicos de grande importância para a região Amazônica, para o Brasil e Mundo.

Foram selecionados, para a discussão, cinco programas/projetos de pesquisas sobre biodiversidade de longa duração, suportados pelo LIS/INPA e com políticas de dados vigentes que objetivam o compartilhamento e integração dos dados gerados. Esses dados são extremamente importantes para a criação, desenvolvimento e evolução de políticas públicas para proteção da biodiversidade e manutenção do crescimento sustentável da região Amazônica. Os projetos/programas, aqui apresentados, foram de grande valia no processo de descoberta de padrões de dados e análise do fluxo de disseminação dos dados sobre a biodiversidade brasileira apresentados nesta tese.

2.6.1 Programa de Grande Escala da Biosfera-Atmosfera na Amazônia – LBA

O Programa de Grande Escala da Biosfera-Atmosfera na Amazônia (LBA) (do inglês: *Large Scale Biosphere-Atmosphere Experiment in Amazonia*) é um programa de pesquisa multidisciplinar, liderado pelo governo brasileiro com projetos de pesquisas financiados por várias agências nacionais (como o MCTIC, o CNPq, a FAPESP, a FINEP/PPG7, etc.) e internacionais (com destaques para a *National Aeronautics and Space Administration* (NASA) e a *National Science Foundation* (NSF), dos EUA, a Comissão Europeia, o Instituto Interamericano de Pesquisas sobre Mudanças Globais, etc.) (LUIZAO; NOBRE; MANZI, 2005; BATISTELLA *et al.*, 2007).

A região Amazônica está sob rápida e intensa transformação relacionada ao processo de desenvolvimento e a ocupação. Por esta razão, o Programa LBA busca compreender como as mudanças, tanto do uso e cobertura da terra, quanto do clima da região, podem afetar os processos biológicos, químicos, físicos e o desenvolvimento sustentável na região, além das interações com o clima regional e global (BATISTELLA *et al.*, 2007). O Programa LBA estuda o funcionamento dos ecossistemas amazônicos, em todas as suas vertentes, e o sistema amazônico como uma entidade regional no sistema Terra, assim como as causas e efeitos das mudanças em curso na Amazônia.

O Programa LBA é coordenado, desde seu início em 1998, pelo MCTIC. Inicialmente, a gerência executiva do Programa LBA esteve a cargo do Instituto Nacional de Pesquisa Espaciais (INPE) e do INPA, tendo a NASA e outras instituições dos Estados Unidos e da Europa como

parceiros. Em setembro de 2007, o LBA tornou-se um programa do governo brasileiro, sob a coordenação científica apenas do INPA.

O Programa LBA é considerado uma das maiores experiências científicas do mundo na área ambiental. Em 2007, o LBA já somava cerca de 156 projetos de pesquisas, sendo que 100 destes já haviam sido finalizados. Esses projetos foram desenvolvidos por 281 instituições nacionais e estrangeiras (Programa LBA, 2007).

O Programa LBA é centrado em duas questões tratadas por meio de pesquisas multidisciplinares que abrangem estudos em ciências físicas, químicas, biológicas e sociais (LUIZAO; NOBRE; MANZI, 2005; BATISTELLA *et al.*, 2007):

1. Como funciona a Amazônia, na forma de um sistema regional, com respeito aos ciclos de água, energia, carbono, gases do efeito estufa e nutrientes?
2. Como as mudanças de uso da terra e do clima podem afetar o funcionamento físico, químico e biológico dos ecossistemas amazônicos?

O Programa LBA passou por diversas fases: A pré-LBA (1996 à 1998); a primeira fase do Programa LBA (1998 à 2007); e a segunda e atual fase começou em 2007.

Na sua primeira fase o Programa LBA foi dividido em sete componentes:

- Física do Clima;
- Dinâmica do Carbono;
- Biogeoquímica;
- Química Atmosférica;
- Hidrologia de Superfície e Química da Água;
- Mudanças de Uso e Cobertura do Solo;
- Dimensões Humanas (Sendo este incorporado a partir de uma recomendação do Comitê Científico do LBA em 2002).

As relações de interdependência de informações se dão em quase todas as direções entre os sete componentes do LBA, como apresentado na Figura 9. As setas apresentam o direcionamento do fluxo dos dados e informações entre os componentes, apontando quem consome dados e quem fornece dados entre os componentes do Programa LBA. Por exemplo, o componente Dinâmica do Carbono provê dados para os componentes: Química da Atmosfera, Biogeoquímica, Hidrologia de Superfície e Química da Água. Dinâmica do Carbono consome dados dos componentes: Física do Clima, Química da Atmosfera, Biogeoquímica, Mudanças do Uso e Cobertura do Solo e Hidrologia de Superfície e Química da Água.

Fonte: Adaptada de Santos e Campos dos Santos (2003).

Figura 9 – Componentes do Programa LBA e suas interligações

O Programa LBA executou as pesquisas no âmbito de cada um desses componentes de forma paralela, integrando o conhecimento por meio de várias medidas:

- Planejamento em conjunto das atividades do programa;
- Compartilhamento de sítios de pesquisa, equipamentos e apoio logístico;
- Desenvolvimento de modelos de simulação acoplando processos-chave da química, física e biologia, atuando em várias escalas temporais e espaciais;
- Síntese conjunta e integração de resultados;
- Intercâmbio de dados e metadados por meio do Sistema de Dados e Informações do Programa LBA (LBA-DIS).

O LBA-DIS é um grupo do LBA responsável por validar, arquivar e distribuir os dados do Programa LBA. A Figura 10 apresenta o fluxo dos dados e metadados gerenciados pelo LBA-DIS, onde:

- **Passo 1:** Os grupos de pesquisas produzem conjuntos de dados em suas pesquisas;
- **Passo 2:** Os pesquisadores líderes dos grupos de pesquisas cadastram os conjuntos de dados, ou apenas os metadados, em uma primeira fase e nas demais fases o conjunto de dados pode ser atualizado até a liberação da sua versão final, que se dá com a documentação

Fonte: Adaptada de [Batistella et al. \(2007\)](#).

Figura 10 – O fluxo de dados no LBA-DIS

completa do conjunto de dados. Os pesquisadores utilizam, nesta fase, o *LBA Metadata Editor* (LME), uma ferramenta para cadastro de metadados do Programa LBA;

- **Passo 3:** Os pesquisadores transferem seus dados, por meio da Web, para o centro de dados do LBA, onde os dados são mantidos;
- **Passo 4:** Os pesquisadores e a comunidade do LBA podem consultar os conjuntos de dados, mantidos pelo LBA-DIS, mediante o sistema de busca Beija-Flor, que utiliza as informações contidas nos metadados para realizar as buscas.

Para facilitar o uso dos dados por pesquisadores, internos e externos, do Programa LBA, cada conjunto de dados é documentado de modo que possa ser útil pelo maior tempo possível, mesmo após a conclusão dos projetos que os geraram. O MCTIC tem o compromisso de manter os dados do LBA *on-line* por pelo menos 20 anos após a conclusão da fase de coleta de dados, garantindo, assim, a perenidade para nova utilização dos dados em pesquisas científicas futuras.

No entanto, apesar dos sistemas de informação adotados pelo LBA, como o Beija-Flor e LME, possuírem mecanismos considerados eficientes em gerência de metadados e no armazenamento de dados, não é garantido, de forma efetiva, a perenidade e a qualidade dos dados para a reutilização dos mesmos a longo prazo. Estas ainda são questões em aberto. As pesquisas e desenvolvimento de novas tecnologias de informação, para atacar essas questões, são dependentes de recursos humanos especializados, que são escassos na região.

Fonte: Adaptada de [Batistella et al. \(2007\)](#).

Figura 11 – Focos de pesquisa do Programa LBA

Na segunda fase do Programa LBA, iniciada em 2007, buscou-se trabalhar, de forma mais integrada, seus domínios disciplinares para ampliar o entendimento sobre o funcionamento dos ecossistemas da região e integrar as dimensões sociais e econômicas às pesquisas ambientais de ponta.

Visando a maior integração dos estudos, foi proposta uma nova reestruturação para os domínios disciplinares do programa. Por conseguinte, o plano científico atual ([BATISTELLA et al., 2007](#)) consolidou as sete áreas iniciais de estudos do Programa LBA em três grandes áreas integradas:

1. Interações físico-químicas multi-escalares na interface biosfera-atmosfera na Amazônia;
2. Dimensões sociais das mudanças ambientais e as dinâmicas de uso e cobertura da terra na Amazônia;
3. Processos físico-químico-biológicos nos sistemas aquáticos e terrestres e suas interações.

Nesta nova fase do Programa LBA torna-se prioritária a criação de mecanismos para a aplicação dos resultados das pesquisas científicas no desenvolvimento sustentável da Amazônia. A Figura 11 apresenta os três domínios propostos, indicando a forte interação em suas áreas de interface.

No contexto desses domínios disciplinares, três focos agregam as principais questões de pesquisa do Programa LBA (na segunda fase):

- O ambiente amazônico em mudança;

- A sustentabilidade dos serviços ambientais e os sistemas de produção terrestres e aquáticos;
- A variabilidade climática e hidrológica e sua dinâmica: respostas, adaptação e mitigação.

Seguindo este caminho, o Programa LBA tem contribuído para:

- Melhorar os modelos de previsão climática;
- Medição das emissões de carbono das hidrelétricas na Amazônia e o potencial uso do metano para geração de energia elétrica adicional nas usinas;
- Realizar novas medidas reais de densidade da madeira no sul da Amazônia, mostrando que biomassa acumulada é menor do que em estimativas anteriores.

Segundo seu plano científico atual ([Programa LBA, 2007](#)), o Programa LBA tem um importante papel na formação de recursos humanos, até 2007 foram formados mais de 500 mestres e doutores brasileiros, e foram publicados mais de 1000 artigos em periódicos especializados, livros e capítulos de livros e mais de 200 artigos em edições especiais dedicadas ao programa.

2.6.2 Programa de Pesquisas em Biodiversidade – PPBio

O Programa de Pesquisas em Biodiversidade foi criado a partir de demandas vindas da sociedade brasileira e elaborado em conformidade com a CDB e com a Política Nacional de Biodiversidade (PNB). A CDB é um tratado das Organizações das Nações Unidas (ONU) que foi estabelecido durante a Conferência das Nações Unidas sobre Meio Ambiente e Desenvolvimento (CNUMAD), conhecida também como ECO-92, realizada na cidade do Rio de Janeiro em junho de 1992 ([CBD, 2013](#)).

O Ministério do Meio Ambiente (MMA) formulou a PNB mediante a realização de estudos no período de 1998 a 2001, enfocando o tema com base nos diferentes aspectos, como ([PNB, 2002](#)):

- A adequação a legislação brasileira com a CDB;
- A síntese sobre a biodiversidade brasileira;
- A comparação da estratégia brasileira com as de outras nações;
- A síntese do conhecimento tradicional associado à biodiversidade no Brasil.

Participaram da elaboração da proposta da PNB o governo federal brasileiro, as organizações não-governamentais (ONG), a comunidade acadêmica, os indígenas, a população e os empresários. Em paralelo a estas atividades, o MMA promoveu uma avaliação dos cinco biomas brasileiro em

quatro reuniões que resultaram no texto final da PNB. Em 22 de agosto de 2002, foi lançado o decreto número 4.339, que estabeleceu a PNB (PNB, 2002).

Nesse contexto, em 2004, o PPBio foi criado com a missão de desenvolver uma política de Ciência, Tecnologia e Inovação aplicada em diversos campos do conhecimento para a gestão, integração e disseminação de informações sobre biodiversidade que possam ser utilizadas para diferentes finalidades (PEIXOTO *et al.*, 2009).

O PPBio tem como principal objetivo o desenvolvimento da capacidade de ampliação da disseminação do conhecimento da biodiversidade brasileira de forma planejada e coordenada, tanto regionalmente como nacionalmente. Para alcançar os seus objetivos, o PPBio conta com quatro eixos de atuação:

- Apoio à implantação e manutenção de redes de inventário da biota brasileira;
- Apoio a modernização de acervos biológicos (coleções *ex situ*);
- Apoio a pesquisa e desenvolvimento em áreas temáticas da Biodiversidade;
- Desenvolvimento de Ações Estratégicas para Políticas de Pesquisa em Biodiversidade.

O PPBio possui um modelo de gestão descentralizado, suas ações e atividades são coordenadas pelas agências de fomento à pesquisa e com apoio de dois institutos de pesquisas e de uma universidade, denominados de Núcleos Executores (NEx).

Na Amazônia o PPBio iniciou suas atividades coordenado por dois NEx. Na parte da Amazônia Oriental, a coordenação é realizada pelo NEx MPEG e na parte da Amazônia Ocidental, pelo NEx INPA. Posteriormente, o PPBio foi expandido para o semi-árido, criando o terceiro NEx, na Universidade Estadual de Feira de Santana (UEFS). O papel fundamental de cada NEx é de coordenar esforços com outras instituições de ensino, pesquisa e gestão sócio-ambiental, denominados como Núcleos Regionais (NRs). Os NRs atuam na geração de conhecimento sobre a biodiversidade das regiões, ou de biomas específicos.

Em outubro de 2009, por meio do lançamento do edital MCT/CNPq/PPBio número 60/2009, foi aprovado a criação de três redes de pesquisa:

- **Amazônia Ocidental** – abrangendo oito projetos de pesquisa e seis instituições, dentre elas: INPA, Universidade Federal do Amazonas (UFAM), Universidade Federal de Rondônia (UNIR), Universidade Federal de Mato Grosso (UFMT), Universidade Federal de Roraima (UFRR) e a Universidade Federal do Acre (UFAC);
- **Amazônia Oriental** – com seis projetos e cinco instituições, dentre elas: MPEG, Universidade Federal do Oeste do Pará (UFOPA), Universidade Federal do Pará (UFPA), Universidade Estadual do Maranhão (UEMA), Universidade do Estado de Mato Grosso (UNEMAT) e a Universidade Federal do Tocantins (UFT);

- **Semi-Árido** – com oito projetos e quatro instituições, dentre elas: UEFS, Universidade Estadual de Santa Cruz (UESC), Universidade Federal de Sergipe (UFS) e Universidade Federal do Rio Grande do Norte (UFRN).

Para o PPBio, um sistema integrado de informações sobre a biodiversidade é essencial ao conhecimento, uso e conservação do patrimônio biológico (OLIVEIRA *et al.*, 2008). Os dados do PPBio são coletados de forma padronizada, ou seja, utilizando o mesmo desenho experimental para permitir uma análise robusta dos padrões encontrados.

A metodologia RAPELD é utilizada pelo PPBio. Esta metodologia é uma combinação de inventários rápidos (RAP - *Rapid Assessment Program*) com Pesquisas Ecológicas de Longa Duração (PELD) (MAGNUSSON *et al.*, 2005). A metodologia RAPELD é baseada no uso de protocolos de amostragem padronizados e integrados espacialmente. Uma descrição preliminar do RAPELD pode ser encontrada em (MAGNUSSON *et al.*, 2005).

Um dos objetivos do PPBio, para o monitoramento da biodiversidade brasileira, é implementar um delineamento padronizado, que permita a comparação entre um determinado grupo biológico em relação as outras regiões, como também, entre os diferentes grupos biológicos em uma mesma região (MAGNUSSON *et al.*, 2005).

O delineamento para inventários do PPBio busca ajustar as necessidades amostrais de diversos grupos biológicos dentro das mesmas unidades amostrais (parcelas), propiciando a integração dos dados bióticos e abióticos, bem como a integração de dados de diferentes grupos biológicos (ZUQUIM; COSTA; PRADO, 2007).

Para o PPBio, dados coletados em terras públicas, ou por funcionários públicos, ou utilizando a infraestrutura fornecida por financiamento público são de propriedade pública e devem ser disponibilizados em tempo hábil para os profissionais que os demandaram. Todos os dados do PPBio tornam-se públicos depois de dois anos após as coletas, conforme a política de disponibilidade de dados do PPBio (DOU, 2009).

O PPBio possui um Núcleo de BioGeoInformática (NBGI), formado por um conjunto de profissionais de TI e biólogos, cujo objetivo é desenvolver aplicativos para o gerenciamento das informações biológicas geradas pelo PPBio (PPBIO, 2013).

A coordenação do NBGI encontra-se dividida entre MPEG e INPA. O NBGI conta, também, com a colaboração de profissionais lotados nos núcleos regionais, principalmente na UFPA e UFT. No MPEG, o grupo trabalha no desenvolvimento de aplicativos como: a customização de ferramentas, como o BRAMS e *Specify*, utilizados para o gerenciamento de coleções biológicas; o desenvolvimento do Sistema de inventário biológicos da Amazônia (SINBIO); desenvolvimento de aplicativos Web para intercâmbio de dados entre coleções científicas e didáticas; construção do catálogo eletrônico das coleções zoológicas e herbário. O NBGI promove e ministra cursos, treinamentos e atividades de capacitação para toda a comunidade do PPBio.

Contudo, o estudo dos componentes do PPBio, que produzem grande volume de dados sobre a biota da Amazônia e o seu gerenciamento, demanda ferramentas automáticas inseridas no contexto da Web Semântica, oferecendo assim suporte para integração e disseminação de resultados. Neste contexto, o LIS objetiva aplicar tecnologias da Web Semântica, como metadados, modelos e ontologias, aos problemas de interoperabilidades de dados e sistemas sobre a biosfera e atmosfera da Amazônia (SANTOS *et al.*, 2011). Além do mais, recuperar informações utilizando um simples mecanismo de busca de palavras chaves em textos na Web não é eficaz, devido ao conjunto de atividades e processos existente na geração de conhecimento científico (SANTOS *et al.*, 2011).

2.6.3 Programa de Pesquisas Ecológicas de Longa Duração – PELD

O Programa de Pesquisas Ecológicas de Longa Duração (PELD) é uma iniciativa pioneira no sentido de obter informações relevantes para a conservação da biodiversidade e uso sustentável dos recursos naturais dos ecossistemas brasileiros.

O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) desde 1999 fomenta a execução do PELD, inserido-o no programa “Ciência, Tecnologia e Inovação para Natureza e Clima” do MCTI (CNPQ, 2013). O PELD é membro do *International Long Term Ecological Research* (ILTER), que reúne cerca de 40 países e suas redes de pesquisas sobre em Ecologia de Longa Duração (ILTER, 2013).

O PELD possui uma rede de sítios de pesquisas, presentes nos diferentes ecossistemas brasileiros, integrados para o desenvolvimento e o acompanhamento de pesquisas ecológicas de longa duração, com objetivo de compreender e prever mudanças ambientais. Atualmente, o PELD conta com uma rede de 30 sítios de pesquisa⁸ (Portal Brasil, 2013). O objetivo central dos sítios de pesquisa do PELD é a investigação dos padrões de funcionamento dos ecossistemas e dos impactos causados pelas perturbações antrópicas e das mudanças globais.

O INPA coordena o Sítio Floresta Amazônica - Manaus do PELD, onde os pesquisadores vinculados ao INPA conduzem pesquisas nos ecossistemas aquáticos e terrestres da floresta tropical úmida. As pesquisas do PELD no INPA abrangem três reservas florestais: a Reserva Florestal Adolpho Ducke, a Reserva Florestal Cuieiras, e a Reservas de Fragmentos Florestais.

Nestas reservas, são encontradas amostras de três níveis de impacto antrópico, que é todo o impacto resultante da ação humana na região, que são:

- (1) o impacto praticamente nulo em florestas primárias;
- (2) o impacto reduzido em áreas de extração seletiva de madeira;
- (3) o impacto alto em fragmentos florestais isolados por pastagens.

O Sítio Floresta Amazônica - Manaus do PELD possui um repositório de dados que

⁸ <http://cnpq.br/sitios-peld/>

armazena e dissemina os dados levantados na reserva florestal Adolpho Ducke, pelo Projeto Dinâmica Biológica de Fragmentos Florestais (PDBFF) na estação experimental de silvicultura florestal, e na Reserva Florestal do Cuieiras.

O SiBBR possui o repositório de dados ecológico, que foi desenvolvido como uma ferramenta para viabilizar o armazenamento quanto a reutilização das bases de dados produzidas por iniciativas como o PELD e PPBio. Os metadados do PELD seguem o padrão EML ([FEGRAUS et al., 2005](#)). Para uma maior abrangência de uso dos dados, o PELD segue a mesma política de dados adotada pelo PPBio.

2.6.4 Projeto Dinâmica Biológica de Fragmentos Florestais – PDBFF

A rápida perda e fragmentação de florestas antigas estão entre as maiores ameaças à biodiversidade tropical ([GARDNER et al., 2009](#)). O PDBFF foi criado devido a um debate científico acalorado, de meados da década de 70, sobre a aplicabilidade da teoria da biogeografia de ilhas para o planejamento de unidades de conservação. Este debate, conhecido internacionalmente pela sigla “SLOSS” (*Single Large Or Several Small reserves of equal area*) procurava avaliar a importância da manutenção de uma reserva florestal grande ou de várias pequenas de igual tamanho ([MACARTHUR, 1967](#)). Naquele tempo, entretanto, os argumentos sobre o “SLOSS” eram mais sobre esta teoria ecológica do que sobre dados reais, já que haviam poucos dados disponíveis, o que incentivou muitos ecólogos a estudar ecossistemas fragmentados ou insulados a partir da década de 70.

Concomitantemente a esta discussão sobre o “SLOSS” foi a criação do Distrito Agropecuário da Superintendência da Zona Franca de Manaus (SUFRAMA), na cidade de Manaus, que permitiu a condução de experimentos em grande escala pelo INPA nos 600 mil hectares destinados basicamente ao uso agrícola e pastoril.

Sob esta oportunidade, Thomas Lovejoy, por meio da *World Wide Fund for Nature* (WWF), associado ao INPA e a SUFRAMA, planejaram a criação de 23 fragmentos florestais, dentro de três grandes fazendas de gado, a serem criados pela SUFRAMA. Assim, em 1979, nasceu o projeto “Tamanho Mínimo Crítico de Ecossistemas” coordenado por Richard Bierregaard. Na década de noventa, o projeto foi incorporado a Coordenação de Pesquisas em Ecologia do INPA por meio do convênio entre o MCTIC/INPA e o *Smithsonian Institution* (SI) dos EUA.

O PDBFF têm cooperação bilateral entre o INPA e o SI. As perspectivas do entendimento da fragmentação do habitat é a justificativa vital para a continuação deste projeto. Além disso, o PDBFF desempenha um papel preponderante na formação de cientistas amazônicos e sustenta a pesquisa de longo prazo sobre os fenômenos das mudanças globais e da regeneração da floresta, além de estudos ecológicos básicos.

Contudo, o PDBFF é um dos projetos de pesquisa ecológica mais duradouras e influentes ainda em existência ([GARDNER et al., 2009](#); [PERES et al., 2010](#)). Assim como o PELD, o

Fonte: Adaptada de [TEAM \(2016\)](#).

Figura 12 – Como o TEAM trabalha

PDBFF utiliza a infraestrutura do SiBBR para disseminação de dados. Os metadados do PDBFF seguem o padrão EML.

2.6.5 Projeto de Ecologia, Avaliação e Monitoramento de Florestas Tropicais – TEAM

O Projeto de Ecologia, Avaliação e Monitoramento de Florestas Tropicais (TEAM - *Tropical Ecology, Assessment and Monitoring*) foi criado em 2002 pela necessidade da criação de uma rede mundial de monitoramento da biodiversidade a longo prazo utilizando protocolos padronizados ([MARTINS; SANDERSON; SILVA-JÚNIOR, 2006](#)).

O objetivo do TEAM é monitorar diversos grupos da fauna, flora e variáveis ambientais ao longo de toda a faixa tropical do planeta, dessa maneira, o TEAM fornece um sistema de alerta sobre o estado da biodiversidade que pode ser direcionados à ações de conservação. Para tanto, a Fundação *Gordon & Betty Moore*⁹ forneceu um financiamento ao Centro para Ciência de Biodiversidade Aplicada (CABS - *Center for Applied Biodiversity Science*) da *Conservation International* (CI) para o início do TEAM ([MARTINS; SANDERSON; SILVA-JÚNIOR, 2006](#)).

O TEAM utiliza uma metodologia padrão para medir e comparar plantas, mamíferos terrestres, aves e o clima em uma série de florestas tropicais, com lugares relativamente intocados em relação aos lugares afetados pelas atividades humanas. Para tanto, o TEAM conta com cerca de dezesseis sítios de pesquisas espalhados pela África, Ásia e América Latina com apoio de cientistas comprometidos com métodos padronizados de coleta de dados para quantificar com as plantas e os animais respondem à pressões externas, como as alterações climáticas e a ocupação humana.

O fluxo de trabalho do TEAM é apresentado na Figura 12:

⁹ <http://www.moore.org>

- Passo 1 – o TEAM mantém sítios de monitoramento ao redor da faixa tropical do planeta para monitorar a Biodiversidade;
- Passo 2 – os dados são coletados de maneira padronizada em cada sítio permitindo a troca de informações;
- Passo 3 – com ajuda de ferramentas móveis os pesquisadores podem contar com mais facilidades durante a coleta de dados no campo;
- Passo 4 – o TEAM armazena os dados provenientes do campo em servidores e banco de dados;
- Passo 5 – os dados são globalmente disseminados em tempo (quase) real por utilizar tecnologias avançadas de disseminação;
- Passo 6 – os dados coletados e disseminados pelo TEAM são utilizados por diversas ONGs, governos, agências, cientistas, educadores, etc. ao redor do mundo através da Web.

O TEAM escolheu trabalhar em estações de pesquisa já estabelecidos em florestas tropicais, conduzidas para a conservação da biodiversidade. O TEAM possui um sítio de pesquisas em Manaus, que teve o início das atividades em 2002, sobre a administração do INPA. A Figura 13 demonstra as três estações pertencentes aos sítios de pesquisas em Manaus, sendo que a primeira estação está localizada na Reserva Florestal Adolfo Ducke, a segunda na saída da BR 174, que liga Manaus a cidade de Boa Vista, e a terceira está localizada no ramal ZF2.

Contudo, o TEAM é um programa com mais de dez anos em funcionamento, uma vez que os seus resultados mostram o valor das coletas em escala global de longo prazo, com dados padronizados e analisados por biólogos e profissionais locais. O TEAM também pode servir como um modelo de programa de acompanhamento para outras partes da CDB. O objetivo da implementação de protocolos padronizados para monitoramento de biodiversidade a nível global poderia ser realizado dessa maneira. O TEAM mantém um banco de dados e metadados disponível para consulta na Web utilizando os padrões de dados e metadados EML e DwC.

2.7 Análise do Fluxo de Dados sobre Biodiversidade das Pesquisas Amazônicas

Nas seções anteriores apresentamos programas e projetos científicos executados na abrangência do LIS/INPA, considerados de extrema importância para o estudo sobre a biodiversidade na Amazônia, além de possuírem experimentos de longa duração sobre pesquisas transdisciplinares.

Com base nestes projetos/programas de pesquisas, foi levantado um estudo de como estes projetos utilizam as tecnologias da IB para a gestão e disseminação de seus dados, dando

Fonte: Elaborada pelo autor.

Figura 13 – Estações do TEAM do sítio de Manaus

um visão macro de como ocorre o fluxo de dados no dia-a-dia das pesquisas de biodiversidade no âmbito da Amazônia.

A Figura 14 apresenta uma visão geral das tecnologias da IB, como os padrões de dados e metadados (seção 2.2), as ferramentas de gestão de dados (seção 2.3), os protocolos de disseminação de dados (seção 2.4) e os repositórios de dados sobre biodiversidade (seção 2.5). A grande nuvem azul representa a abrangência da internet. Os retângulos em cinza, dentro da nuvem da internet, correspondem as ferramentas e repositórios de disseminação de dados, que são baseados em tecnologias Web. Os retângulos em cinza, dentro da nuvem da internet, representam as repositórios de dados na Web, utilizados para disseminação de dados sobre biodiversidade. Os retângulos em brancos representam os protocolos de disseminação de dados sobre biodiversidade. Os silos verdes representam os SGBD utilizados por ferramentas de gestão de dados sobre biodiversidade, como o Specify (retângulo vermelho). Os ícones de arquivo, em amarelo, representam os padrões de dados e metadados. Os retângulos em vermelho, fora da nuvem da internet, representam as ferramentas *standalone* utilizadas para administrar dados

Fonte: Elaborada pelo autor.

Figura 14 – Fontes de dados sobre biodiversidade na Web

sobre biodiversidade.

A Figura 15 apresenta uma visão geral das ferramentas agrupadas com os respectivos programas e projetos científicos (seção 2.6). Cada ferramenta de gestão de dados de biodiversidade, representada pelo ícone e descrição, apontam para os seus respectivos padrões de dados e metadados utilizados para disseminação de dados sobre biodiversidade.

O sistema Beija-flor, adotado pelo Programa LBA, dissemina dados heterogêneos produzidos pelos experimentos do Programa LBA e tem como principal característica a busca com base nos metadados, que são gerenciados pelos pesquisadores usando a ferramenta LME, os dados são disponibilizados nos formatos DwC, EML, CSV e XML.

O Metacat é utilizado pelo PDBFF, PPBio e PELD, abordado na seção 2.3.3.

O Morpho (seção 2.3.4) é um editor de EML que se conecta ao Metacat para armazenar e realizar buscas nos dados armazenados.

O Specify (seção 2.3.2) dissemina dados de coleções do PPBio por meio de sua versão Web.

O BRAHMS (seção 2.3.1), em sua interface Web, exporta dados das consultas do usuários por meio de arquivos de planilhas eletrônicas.

O speciesLink (seção 2.5.2), por meio da sua interface Web, permite os usuários realiza-

Fonte: Elaborada pelo autor.

Figura 15 – Infraestrutura atual para disseminação de dados sobre biodiversidade dos programas e projetos científicos do INPA e MPEG

rem consultas sobre os dados da rede e exportar os conjuntos de dados em arquivos de planilhas eletrônicas.

Todas as ferramentas permitem que os usuários realizem consultas, baseadas nas informações dos metadados e dos padrões de dados, por meio da Web, porém a abrangência dos resultados das consultas fica limitada aos dados disponíveis em cada ferramenta, não existe uma consulta integrada nos dados disponíveis nestes repositórios, nem sequer formas de consultas integradas com outras fontes de dados científicos.

As consultas podem ser exportadas pelos usuários em arquivos de planilhas eletrônicas e metadados no formato XML. Basicamente, quase todos os dados sobre biodiversidade são disseminados em formato de planilhas eletrônicas, ou por arquivos de texto tabulados, associados à arquivos XML, contendo informações dos seus metadados.

No entanto, estes formatos de dados não permitem uma rápida colaboração com outras áreas de conhecimento, para uma abrangente descoberta de informações entre os conjuntos de dados dispersos e heterogêneos. Ademais, quando os cientistas e os não-cientistas descobrem, acessam e utilizam dados provenientes de fontes desconhecidas, então, são forçados a aprender detalhes sobre o esquema de dados, esquemas de nomes e de termos que outras pessoas ou grupos criaram, bem como a sintaxe dos dados, para então poder trabalhar na busca de informações relevantes a suas pesquisas. Outro agravante é o fato que a maioria dos biólogos não possuem expertise em técnicas e ferramentas computacionais para explorar, com mais profundidade, os dados e encontrar correlações nos fatos entre fontes distintas de informações sobre biodiversidade.

2.8 Considerações Finais

Segundo Hobern, os governos ao redor do mundo não conseguiram atingir a meta de redução da perda de biodiversidade prevista para 2010 (HOBERN *et al.*, 2013). Um dos fatores-chaves foi a falta de informação, que deixou os governos impossibilitados de tomar melhores decisões e de criar novas políticas públicas que visassem à conservação da biodiversidade.

Nesse sentido, a IB atua como ponto fundamental para alcançar a redução da perda da biodiversidade mundial, por meio da mobilização de todos os dados mundiais (antigos e novos) sobre biodiversidade de forma estruturada e padronizada. A partir deste ponto será possível criar uma vasta gama de utilização dos dados sobre biodiversidade, além de novas oportunidades para pesquisas e fortalecer a formulação de políticas, relacionadas à conservação da biodiversidade.

Neste capítulo foram apresentados os principais conceitos relacionados com a padronização de dados e metadados sobre biodiversidade, ferramentas de gestão de dados, protocolos de disseminação de dados e repositórios de dados além de destacar alguns projetos/programas suportados pelos nosso parceiro de pesquisa que guiou à análise do fluxo de disseminação de dados sobre a biodiversidade da Amazônia, sendo muito útil para o embasamento desta tese.

Os padrões de dados e metadados sobre biodiversidade foram desenvolvidos com a finalidade de facilitar o intercâmbio de informações e permitir a criação de redes de compartilhamento de dados sobre biodiversidade pela Web.

Os provedores de informações sobre biodiversidade (coleções, museus, institutos, universidades, projetos de pesquisas, etc.) administram seus dados de diversas maneiras (por meio de sistemas de informação, banco de dados, arquivos de planilhas eletrônicas, etc.).

Para disseminar estes dados via Web, foi necessário criar protocolos (softwares) que mapeassem as diferentes estruturas de dados, existentes nos bancos de dados sobre biodiversidade, para um padrão de dados e metadados em comum na comunidade de IB. Os principais protocolos de disseminação da IB foram discutidos na seção 2.2, que auxiliam no funcionamento das redes de compartilhamento de dados de biodiversidade, como o *speciesLink* e GBIF.

Todas as tecnologias da IB citadas neste capítulo são importantes para o desenvolvimento da infraestrutura semântica proposta.

Os programas/projetos mencionados na seção 2.6 formam uma vasta fonte de conhecimento sobre a biodiversidade da Amazônia. Os dados produzidos por esses programas/projetos formam a base de conhecimento para diversas medidas de proteção e entendimento dos complexos ecossistemas amazônicos.

No entanto, apesar desses programas/projetos terem políticas de dados vigentes, que objetivam o compartilhamento e integração dos dados gerados, os seus dados sobre biodiversidade, disponíveis na Web, estão em formatos que não permitem uma rápida colaboração com outras áreas de conhecimento, tanto para descobertas de novos conhecimentos, a partir da análise dos

relacionamentos entre as entidades e compreensão transdisciplinar dos fatos que as descrevem, como, também, para serem reutilizados e relacionados com dados de outras áreas do conhecimento. Ademais, existem muitos dados disponíveis sobre um determinado recurso, como um táxon, mas todo esse conhecimento encontra-se disperso e desconectado em diversos repositórios de dados com diferentes formatos de dados e metadados. Os usuários destes dados necessitam de mecanismos para realizar a integração. Até então, não existiam ferramentas de IB para realizar uma integração desses dados de forma fácil aos usuários.

Até então, muitos dados disponíveis na Web são utilizados apenas para a apresentação de conteúdo, em páginas Web por exemplo, deixando a interpretação das informações no entendimento humano.

A Web Semântica surge neste contexto como uma nova geração de representação de informações, mais também para a automação, integração, interoperabilidade e reutilização de recursos. As tecnologias da Web Semântica oferece uma oportunidade de adicionar conteúdo semântico aos dados de biodiversidade, permitindo que computadores possam inferir novas informações implícitas, acerca dos dados, em um ambiente que permite a integração de diversas outras áreas do conhecimento.

Com base nesse cenário, é proposta a elaboração de uma infraestrutura semântica, aplicando tecnologias da Web Semântica, para integrar os dados sobre biodiversidade de forma fácil. Assim, pode-se demonstrar o grau de transdisciplinaridade das informações contidas nos dados, melhorando o reúso, adoção de padrões e a diminuição dos esforços repetitivos na realização de novas pesquisas.

WEB SEMÂNTICA: TECNOLOGIAS DE INTEGRAÇÃO DE DADOS

3.1 Considerações Iniciais

A internet surgiu como uma forma para interligar computadores em rede, onde mensagens e informações começaram a ser trocas entre os computadores ao redor do mundo. Uma revolução ocorreu no âmbito da internet com o surgimento da *World Wide Web* (WWW) (BERNERS-LEE, 1992; BERNERS-LEE *et al.*, 1993), ou simplesmente Web. Esta revolução permitiu que documentos sejam interligados uns aos outros, não importando em que computador eles estejam na internet, compondo desta maneira a Web. Pessoas comuns começaram a utilizar a Web para navegar, por meio das interligações entre os documentos, procurando por informações relevantes.

Atualmente, uma nova revolução está ocorrendo na Web, no que diz respeito aos conteúdos presentes nos documentos. Os conteúdos estão relacionados a diferentes tipos de assuntos, como: uma lista de contatos, uma lista de preferências de usuário, uma agenda de compromissos ou até uma lista de espécies em extinção. A Web Semântica deve construir uma forma de agregar significado estruturado para o conteúdo dos documentos disponíveis na Web, criando um ambiente onde agentes de software, ao percorrerem página por página dos documentos, possam facilmente realizar tarefas para os usuários (BERNERS-LEE; HENDLER; LASSILA, 2001).

O termo Web Semântica se popularizou em 2001, por meio de uma publicação na revista *Scientific America* intitulada “*The Web Semantic: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*” (BERNERS-LEE; HENDLER; LASSILA, 2001). O trabalho conceitua que a Web Semântica é uma representação capaz de associar significados explícitos aos conteúdos dos documentos disponíveis na internet, sendo que sua principal meta é possibilitar que programas processem e interpretem automaticamente esses documentos. Para DuCharme (2011), a Web Semântica é um conjunto de padrões e melhores

práticas para compartilhar dados e semântica de dados via Web.

O *World Wide Web Consortium* (W3C) é a principal organização de padronização da Web. A missão do W3C é levar a Web ao seu máximo potencial, desenvolvendo protocolos e diretrizes que garantam crescimento a longo prazo (W3C, 2013a).

A Web atual, definida por muitos como a “*Web Sintática*” ou “*Web de Documentos*”, é composta por páginas Web desenvolvidas apenas para a apresentação da informação, logo a estruturação, semântica e o significado dos dados apresentados nessas páginas ficam a cargo dos usuários, que estabelecem o papel de compreender e entender o conhecimento ali apresentado.

Além da “*Web de Documentos*”, o W3C está ajudando, também, a construir tecnologias para apoiar a WD. O objetivo desta WD é permitir que computadores façam o trabalho de forma mais útil, por meio de sistemas que possam apoiar interações confiáveis pela Web (W3C, 2013a).

A liberdade que a Web trouxe aos seus usuários, permitindo a publicação de conteúdos próprios, também é apoiada pela Web Semântica, ou seja, em vez de impor um modelo único para organização dos dados, a Web Semântica permite o surgimento de vários modelos de dados em paralelo, podendo ou não estarem relacionados entre si (POLLOCK, 2009).

Este capítulo aborda as principais tecnologias da Web Semântica que fazem parte do escopo desse trabalho. A Seção 3.2 apresenta a arquitetura em camadas do funcionamento da Web Semântica apresentada pelos seus autores. Os conceitos e recomendações do LD são apresentados na seção 3.3. O resultado de uma pesquisa sobre ferramentas de conversão de arquivos tabulados para o modelo RDF é apresentado na seção 3.4.

3.2 Arquitetura da Web Semântica

Tim Berners-Lee propôs um modelo em camadas para o funcionamento da Web Semântica, conhecido como “*bolo de noiva*” (BREITMAN, 2006). A Figura 16 apresenta o modelo, demonstrando como as principais camadas de tecnologias da Web Semântica estão dispostas.

O modelo em camadas tem como idéia central adicionar gradativamente novas funcionalidades, como o aumento de expressividade, a possibilidade de realizar inferência e a autenticação. Essa arquitetura consiste em construir algo novo em cima do que já existe e não reinventar tudo de novo a cada nova tecnologia (BREITMAN, 2006).

No modelo bolo de noiva (Figura 16) de Tim Berners-Lee, o *Uniform Resource Identifier* (URI) é a base para representação de recursos e conceitos, ou seja, tudo na Web Semântica é representado como uma URI. O RDF é o modelo para representação de dados, formado por um tripla com sujeito, predicado e objeto. Tanto o sujeito e predicado devem ser representados por uma URI e o objeto deve ser uma URI ou um literal. Um conjunto de triplas forma um gráfico de informações. O *Resource Description Framework – Schema* (RFD–S) e a OWL são as bases da Web Semântica, as definições lógicas como classes, propriedades, indivíduos, etc. são definidas

Fonte: Adaptada de Pollock (2009).

Figura 16 – Arquitetura em camadas da Web Semântica

com estas linguagens, que são baseadas no RDF.

A Web Semântica permite, também, o uso de regras para aumentar a expressividade nas definições lógicas. O SPARQL é a linguagem de consulta na Web Semântica. As demais camadas ainda estão em fase de amadurecimento e desenvolvimento.

O RDF, RFD-S, OWL, SPARQL são padrões oficialmente recomendados pelo W3C. Estas tecnologias são consideradas estáveis desde as suas primeiras versões.

Na ciência, os avanços no uso de vocabulários controlados (e ontologias) fornecerão termos bem definidos para preencher as entradas de dados e permitir que os computadores possam ajudar, com mais precisão, os pesquisadores na localização e processamento de dados de interesse (REICHMAN; JONES; SCHILDHAUER, 2011). No entanto, sistemas de metadados atuais fornecem entradas de dados rotuladas para serem preenchidas usando linguagem natural e, portanto, não passíveis de interpretação automática por computadores (REICHMAN; JONES; SCHILDHAUER, 2011).

3.2.1 Uniform Resource Identifier – URI

O URI é uma cadeia de caracteres (*string*) utilizada para identificação de um recurso, que pode ser um página escrita em *Hypertext Markup Language* (HTML), uma imagem, um arquivo, etc. Todo recurso na internet deve ter um identificador único para ser encontrado. Recursos de rede necessitam informar o seu protocolo. A Web utiliza o protocolo *Hypertext Transfer Protocol* (HTTP).

Na Figura 17 é apresentada a *string* de uma URI. A parte da *string* “http://”, dentro do retângulo azul, identifica o protocolo, neste caso o HTTP. A *substring* “www.usp.br”, dentro do

Fonte: Elaborada pelo autor.

Figura 17 – Exemplo de URI

retângulo verde, identifica o nome do *host* que será acessado pela internet. Por fim, a *substring* “/index.html”, dentro do retângulo vermelho, identifica o recurso que está no *host*, neste caso um arquivo HTML.

O *Uniform Resource Locator* (URL) é um tipo de URI. O URL refere-se ao local (*host*) e a um determinado recurso contido nele. A ideia no URL é associar um endereço remoto (nome do *host*) a um nome de recurso que está no *host*.

O *Uniform Resource Name* (URN) é um tipo de URI. O URN é o nome do recurso que será acessado. Geralmente não é informado o protocolo e nem o *host* que se encontra o recurso. Por exemplo, ao se referir a um livro podemos informar apenas o ISBN do mesmo (exemplo: urn:isbn:0-486-27557-4), no entanto não há como saber onde encontra-se este recurso na rede, apenas têm-se o nome do mesmo.

3.2.2 Resource Description Framework – RDF

O *Resource Description Framework* (RDF) é uma recomendação oficial do W3C de 10 fevereiro de 2004 (W3C, 2004b). O RDF é uma linguagem que tem o propósito de representar informações sobre os recursos da Web (URI). Um recurso pode representar qualquer objeto do mundo real, como um livro, uma pessoa ou um local. O RFD é considerado a base da arquitetura de camadas da Web Semântica, como apresentado na Figura 16.

O RDF não é considerado um formato de dados, mas um modelo de dados com sintaxes que permitem o armazenamento em arquivos (DUCHARME, 2011). Este modelo de dados pode expressar fatos, com declarações compostas por três partes, conhecidas também como triplas. Cada tripla é como uma pequena frase que afirma um fato sobre um recurso. A tripla é composta por um Sujeito (S), um Predicado (P) e um Objeto (O). A Figura 18 ilustra uma sentença, ou afirmação, utilizando o modelo de dados RDF.

O Sujeito é um recurso que está sendo descrito pela tripla RDF. O Sujeito deve ser um URI, que aponte para uma página, parte de uma página, um documento XML, um site inteiro ou até mesmo a recurso utilizado para a representação de objeto do mundo real, como um artigo científico, um lugar, um país, etc.

Fonte: Elaborada pelo autor.

Figura 18 – Representação de uma tripla em RDF

Fonte: Adaptada de Pollock (2009).

Figura 19 – Um grafo RDF após a adição de recursos usando FOAF

O Predicado deve ser um URI utilizado para determinar a identificar a propriedade do Sujeito. O Predicado pode ser utilizado para descrever algum aspecto específico, atributo, característica ou relação com outro Sujeito.

O Objeto pode ser um valor literal ou um URI de um outro Sujeito. Os valores literais podem, também, ser representações tipos de dados (DataType) com que esses valores devem ser representado, como número, data, texto, etc.

A Figura 19 apresenta um conjunto de triplas, também conhecido como um gráfico RDF. Os Sujeitos das triplas são representados pelas elipses em cinza, os Predicados são representados pelas setas e os Objetos são representados pelos retângulos. Nota-se que os Sujeitos e os Predicados são sempre um URI. Existe quatro Objetos, sendo três literais e um URI (<http://me.jtpollock.us/foaf.rdf#me>), que também é um Sujeito de uma outra tripla.

Cada declaração em RDF é considerada uma tripla, e um grafo RDF é um conjunto de declarações em RDF, como ilustrado na Figura 19.

O RDF em si é uma linguagem simples e capaz de fazer relacionamentos entre informações.

Essa linguagem descreve as relações entre objetos em termos de propriedades e valores, porém não define mecanismos para descrever tais propriedades e relações entre estas propriedades e outros recursos.

O RFD-S foi criado pelo W3C para fornecer um vocabulário de modelagem de dados para dados em RDF. RDF-S é uma extensão do vocabulário RDF básico. O RDF-S fornece um

sistema de tipos para o RDF similar, em alguns aspectos, a programação orientada a objetos. Os recursos, em RDF-S, podem ser definidos como instâncias de uma ou mais classes. Além disso, o RDF-S permite que as classes sejam organizadas de forma hierárquica, como por exemplo a classe “cachorro” pode ser uma subclasse de “mamífero” que é uma subclasse de “animal”, logo uma instância da classe “cachorro” é, também, implicitamente uma instância da classe “animal”.

Além de descrever as classes de coisas que poderão ser utilizadas para descrever recursos, o RFD-S também descreve as propriedades específicas que essas classes podem ter.

3.2.3 Ontologias e a Web Ontology Language – OWL

A literatura apresenta diversas definições para ontologias, diversos tipos de ontologias, diferentes propostas para aplicação de ontologias (em diferentes áreas de conhecimento) e métodos para a construção de ontologias (metodologias, ferramentas e linguagens). Ontologias têm sido aplicadas em diversas áreas do conhecimento, como Inteligência Artificial, Engenharia de Software, Web Semântica, possibilitando a criação de modelos conceituais claros, concisos e não ambíguos (ALBUQUERQUE, 2016).

Uma ontologia é especificada por meio de seus componentes básicos, como classes, relações, axiomas e instâncias. A ontologia deve ser expressa por meio de uma linguagem padrão (ALMEIDA; BAX, 2003).

Na área de Web Semântica, as ontologias são modelos conceituais, que capturam e explicitam o vocabulário utilizado nas aplicações semânticas. As ontologias servem como base para garantir uma comunicação livre de ambiguidades. As ontologias são consideradas a língua franca da Web Semântica (BREITMAN, 2006).

Embora o RDF e o RDF-S sejam uma boa solução para estruturar conhecimento na Web, são tecnologias bastante limitadas, por exemplo, o RDF-S não é capaz de estabelecer restrições de cardinalidade para propriedades das classes.

No final da década de 90, a necessidade de uma linguagem de descrição de ontologias, com maior poder de expressividade, aliada ao nascimento da Web Semântica, surgiram muitas propostas para uma linguagem padrão de definição de ontologias na Web, como o *Simple HTML Ontological Extensions* (SHOE), a *Ontology Inference Layer* (OIL) e a DAML+OIL (HORROCKS, 2008).

Em 2001, reconhecendo que uma linguagem para ontologias é um requisito prévio para o desenvolvimento da Web Semântica, o W3C criou um grupo de trabalho para o desenvolvimento da *Web Ontology Language* (OWL). O resultado desta iniciativa foi a formação do *Web Ontology Working Group* do W3C. Em 2004 a linguagem OWL foi lançada como a linguagem padrão da W3C para definição de ontologias.

A linguagem OWL foi projetada para ser utilizada por aplicações que precisam processar

o conteúdo da informação, em vez de apenas apresentar informações para os seres humanos.

A OWL tem maior facilidade na interpretação por máquina dos conteúdos da Web em relação a outras linguagens, como XML, RDF e RDF-S, fornecendo um vocabulário adicional juntamente com uma semântica formal.

Diversas ontologias em OWL foram desenvolvidas em várias áreas, como *e-Science*, medicina, biologia, geografia, astronomia, defesa e nas indústrias automobilística e aeronáutica (GRAU *et al.*, 2008).

Apesar do sucesso da OWL, algumas deficiências na linguagem foram apontadas pela sua vasta comunidade de usuários e pelos desenvolvedores de ferramentas OWL e *reasoner*. Em resposta aos comentários e requisições dos usuários da OWL, surgiu a OWL 1.1, uma revisão incremental da linguagem. Mais tarde, extensivas discussões resultaram na OWL 2, uma recomendação da W3C de dezembro de 2012 (W3C, 2012). A OWL 2 é considerada um passo substancial para a evolução da linguagem (GRAU *et al.*, 2008).

A OWL é baseada em lógica de descrição (*Description Logic*). Uma ontologia em OWL é constituída por um conjunto de entidades (*entity*), que necessariamente devem conter um identificador único, um URI ou IRI. As entidades de um ontologia podem ser categorizadas em classes (*Class*), indivíduo (*Individual*), propriedade de relação (*ObjectProperty*), propriedade de dados (*DataProperty*) e as propriedades de anotação (*AnnotationProperty*). Uma classe identifica um conjunto de elementos reais de um determinado domínio de conhecimento, que são chamados de indivíduos, ou instâncias das classes. Uma propriedade de relação permite o relacionamento entre indivíduos e uma propriedade de dados serve para relacionar um valor literal a um indivíduo. Uma propriedade de anotação pode ser aplicada a todas as entidades da ontologia, servindo como notas sobre as mesmas, como rótulos, comentários e observações. Estas anotações servem apenas para o entendimento humano, não são computáveis nos processos de *reasoning*, por exemplo. Ontologias descritas em OWL podem ser usadas em conjunto com informações descritas em RDF (W3C, 2012).

A OWL possui perfis, ou subconjuntos da linguagem. Um perfil, neste contexto, consiste em um versão mais reduzida da OWL. Os perfis OWL permitem estabelecer compromissos em relação ao grau máximo de formalidade (ou expressividade), com consequência nos tipos de inferências possíveis de realizar, e o desempenho pretendido ao nível do processo de inferência.

A OWL possuía na sua primeira versão três perfis, divididos de acordo com a sua expressividade: OWL Lite, OWL DL e OWL full (W3C, 2004a). Com a versão 2 da OWL, passou a ter três perfis subconjuntos da OWL DL, o OWL 2 EL, OWL 2 QL e OWL 2 RL, além do OWL full, como apresentado na Figura 20.

OWL full – perfil que permite o uso de todos os elementos da linguagem, sem qualquer tipo de restrição. Por exemplo, neste perfil um Indivíduo pode ser, ao mesmo tempo, uma Classe.

OWL DL – perfil que também permite o uso de todos os elementos da linguagem, porém

Fonte: Elaborada pelo autor.

Figura 20 – Perfis atuais da Linguagem OWL

com restrições de como estes elementos poderão ser utilizados de forma a garantir a existência de um algoritmo decidível. Por exemplo, neste perfil não é permitido que um Indivíduo seja, ao mesmo tempo, uma Classe.

OWL 2 EL – perfil baseado na família EL++ de lógica descritiva. Este perfil é útil em aplicações que empregam ontologias que declaram um grande número de propriedades ou classes, para o qual a consistência da ontologia, como a *subsumption* da expressão da classe e a verificação da instância, podem ser decididas em tempo polinomial. Este perfil, por exemplo, fornece construtores de classes suficientes para expressar ontologias biomédicas, como o SNOMED CT¹.

OWL 2 QL – perfil baseado na família *DL-Lite* de lógica de descrição. Esse perfil foi criado para permitir o raciocínio computacional (*reasoning*) eficiente com grandes quantidades de dados estruturados de acordo com esquemas relativamente simples. Ele fornece a maioria dos recursos necessários para capturar modelos conceituais, tais como diagramas de classe na *Unified Modeling Language* (UML), diagramas de Entidade de Relacionamento (ER) e esquemas de bancos de dados. Este perfil foi projetado para que dados (asserções) armazenadas em um sistema de banco de dados possa ser consultados por meio de uma ontologia, utilizando um simples mecanismos de reescrita de consultas, onde o SGBD responda sem realizar qualquer alteração nos dados.

OWL 2 RL – perfil voltado para aplicações que exigem raciocínio lógico escalável sem abrir mão de muita expressividade. Este perfil é destinado para ontologias em OWL 2, que podem trocar a expressividade total da linguagem por escalabilidade, e para ontologias em RDF-S, que necessitem de expressividade adicional da OWL 2. Este perfil define um subconjunto sintático da OWL 2 que favorece a implementação utilizando tecnologias baseadas em regras.

Contudo, assim como as outras tecnologias da Web Semântica como o RDF e RFD-S, a OWL tem se mostrado uma tecnologia madura e evoluída. Neste trabalho serão utilizadas ontologias desenvolvidas em OWL para mapear dados sobre biodiversidade vinculados entre si e com outros dados na WD.

¹ <http://www.snomed.org/snomed-ct/>

3.2.4 SPARQL Protocol and RDF Query Language – SPARQL

O *SPARQL Protocol and RDF Query Language* (SPARQL) é uma linguagem e protocolo para realizar consultas e manipulações de triplas RDF na Web Semântica. O SPARQL é um padrão que foi desenvolvido pelo *RDF Data Access Working Group* do W3C. Em janeiro de 2008 o SPARQL 1.0 tornou-se uma recomendação oficial da W3C (W3C, 2008). Em março de 2013 foi lançada a versão 1.1 desta recomendação (W3C, 2013b).

O SPARQL é uma tecnologia muito importante para o desenvolvimento de aplicações na Web Semântica. Assim como a *Structured Query Language* (SQL) fornece uma linguagem padrão para consultas em SGBD, a linguagem SPARQL fornece uma linguagem padrão para consultas em grafos RDF (SEGARAN; EVANS; TAYLOR, 2009).

Na versão 1.0, o SPARQL agregava três documentos em sua especificação:

- **SPARQL Linguagem de consulta para RDF** – documento que abrange a sintaxe para a consulta dos dados na Web Semântica;
- **SPARQL Protocolo para RDF** – documento que especifica como um programa deve passar as consultas SPARQL para um serviço de processamento de consulta SPARQL (*endpoint*) e como esse serviço deverá retornar os resultados obtidos pela consulta;
- **SPARQL Formato da resposta de consulta em XML** – documento que especifica um formato simples, em XML, para ser utilizado como respostas das consultas pelos serviços de processamento de consulta SPARQL.

A versão 1.1. do SPARQL adicionou novos documentos técnicos às recomendações da linguagem e protocolo. As principais mudanças estão relacionadas com a implementação do SPARQL e não tanto com o padrão de consulta dos dados. Foi adicionado padrões para as requisições de atualização, que agora podem ser executadas nos conjuntos de dados. As principais mudanças desta versão são:

- **A extensão de federação** – documento descreve como uma única consulta pode recuperar dados a partir de múltiplas fontes (consultas federadas);
- **A atualização** – documento que especifica a principal diferença entre as versões 1.0 e 1.1 do SPARQL, onde o SPARQL passou de uma linguagem meramente de consulta para uma linguagem capaz de modificar os dados nos grafos RDF, ou seja, esta linguagem permite, além das consultas, a adição, edição e exclusão de triplas em RDF;
- **A descrição do serviço** – documento que descreve como um programa cliente pode perguntar a um motor SPARQL quais recursos ele suporta;

- **O formato JSON para resultados de consulta** – documento que descreve o formato de resultado de consulta baseado na *JavaScript Object Notation* (JSON), equivalente ao formato de consulta em XML do SPARQL;
- **O protocolo HTTP de armazenamento de grafo** – documento que descreve como o protocolo SPARQL, por meio de API, estende a tecnologia *Representational state transfer* (REST) para realizar a comunicação entre um cliente e um processador de consultas SPARQL (*endpoint*).

O SPARQL é fundamental para o desenvolvimento da infraestrutura semântica proposta. Por meio de consultas em SPARQL, juntamente com o poder de expressividade das ontologias, pode-se recuperar informações implícitas dentro de uma base de conhecimento. Além de poder integrar os dados usando as tecnologias da Web Semântica, com SPARQL é possível dar mais sentidos aos dados e fazer com que o computador raciocine e retorne aos usuários mais informações relevantes.

3.2.5 GeoSPARQL

Enquanto, na Web Semântica, o SPARQL é uma linguagem útil para realizar consultar em relações que são explicitamente representadas nos dados (triplas), relações implícitas (inferidas), tais como relações geoespaciais, não podem ser facilmente consultadas apenas utilizando a linguagem SPARQL. O raciocínio geoespacial é essencial em um grande número de domínios, como: hidrologia, respostas de emergências, planejamento de transportes, uso da terra, biodiversidade, etc.

Por exemplo, dados de biodiversidade podem estar associados a nomes de localidades, pontos geográficos, etc. As consultas na área de biodiversidade podem ser do tipo “quantos espécimes foram coletadas ao redor de 3 Km de uma certa localidade”. Realizar esse tipo de consulta vai depender de como os dados são declarados e vinculados a essa localidade e se as relações são explícitas.

No entanto, vincular esses dados com base em relações não declaradas (via inferência na maioria das vezes) não é uma atividade trivial. Os usuários desses domínios há muito tempo vem utilizando SGBD com extensões espaciais para responder a consultas geográficas (EGENHOFER, 1994). Com o crescimento e maturidade destas tecnologias e a tentativa de compartilhar e vincular informações de diferentes repositórios surge a necessidade de interoperabilidade de dados. Com a popularização da Web Semântica e a utilização de soluções baseadas em RDF para realizar relacionamentos e inferências geoespaciais entre bancos de dados, foi possível sanar tais necessidades.

Porém, como vários grupos criaram diversas soluções com fins variados para esse tipo de

tarefa, notou-se a necessidade do estabelecimento de um padrão. O *Open Geospatial Consortium*² (OGC), uma organização sem fins lucrativos centrada na padronização de dados geoespaciais, desenvolveu o padrão GeoSPARQL, com a intenção de unificar o acesso aos dados geoespaciais na Web Semântica (Open Geospatial Consortium, 2012).

O GeoSPARQL é o fruto de uma quantidade significativa de trabalhos anteriores de combinação do RDF e OWL com dados geoespaciais (BATTLE; KOLAS, 2012). Uma lista detalhada e comparativa de abordagens e implementações de suporte a dados geoespaciais na Web Semântica é apresentada em (BATTLE; KOLAS, 2012).

De modo geral, o GeoSPARQL busca resolver problemas com implementações distintas e incompatíveis para a representação e consulta de dados espaciais na Web Semântica. A especificação GeoSPARQL³ contém três componentes principais:

1. A definição de um vocabulário para representar recursos, geometria e relacionamentos;
2. Um conjunto de domínio específico de funções espaciais para ser usado nas consultas SPARQL;
3. Um conjunto de regras de transformação de consultas.

GeoSPARQL define uma ontologia para representar recursos, geometrias, predicados e funções, além de fornecer vários conjuntos de terminologia para especificar relacionamentos topológicos entre geometrias [15].

A Ontologia GeoSPARQL possui duas classes *geo:Feature* e *geo:Geometry* que são subclasse de *geo:SpatialObject*. Estas classes podem ser estendidas para outras ontologias de outros domínios. Recursos que possuem geometria, como uma reserva natural, podem se associar, por meio da propriedade *geo:hasGeometry* com a sua respectiva localização no mundo real, assim poderá ser medida e avaliada. A representação de uma localização do mundo real torna-se um *geo:Geometry*, podendo ser apenas um ponto geográfico a um polígono detalhado que representa o contorno de uma reserva natural, por exemplo. O GeoSPARQL possui duas formas de representação geográficas literais, o *Well-Known Text* (WKT) e o *Geography Markup Language* (GML) por meio das propriedades *geo:asWKT* e *geo:asGML*, respectivamente.

Relações topológicas são suportadas pelo GeoSPARQL de forma padronizada, disponibilizadas sob a forma de propriedades binárias entre as entidades, por meio de funções de filtros geoespaciais (utilizado nas consultas). As relações topológicas suportadas de recursos simples são: *equals*, *disjoint*, *intersects*, *touches*, *within*, *contains*, *overlaps* e *crosses*.

Regras de transformações de consultas SPARQL/GeoSPARQL permitem adicionar abstração nas consultas. O GeoSPARQL, por meio da propriedade *geo:defaultGeometry*, permite

² <http://www.opengeospatial.org>

³ <http://www.opengeospatial.org/standards/geosparql>

declarar relações topológicas e consultar por meio de regras de reescrita de consulta. Se por acaso um *geo:Feature* é usado como Sujeito ou Objeto de uma consulta, esta é automaticamente reescrita para comparar os *geo:Geometry* ligados como padrão, já que um *geo:Feature* é objeto conceitual que existe no mundo real num determinado lugar que deverá estar ligado a um *geo:defaultGeometry* por padrão.

Contudo, dados sobre biodiversidade são ricos em informações espaciais, como dados de ocorrência de espécies. Realizar consultas relacionadas ao raciocínio espacial é um recurso atraente para os usuários de dados de biodiversidade. A infraestrutura proposta pretende explorar o GeoSPARQL para responder perguntas complexas relacionadas a localização dos registros.

3.2.6 Triplestores

O modelo RDF pode ser armazenado em arquivos de texto, seguindo padrões de serialização do RDF, como RDF/XML, o *Notation3* (ou N3), o *Turtle*, o *NTriples* ou o RDF/JSON. Porém, quando é necessário o armazenamento e gerenciamento de grandes quantidades de triplas RDF, os arquivos de serialização de RDF tendem a não atender, de forma eficiente e segura, estas necessidades. Para estes casos, existem outras formas de gerenciamento e armazenamento de grandes volumes de triplas RDF, como o uso de SGBD (como MySQL, PostgreSQL e Oracle) e por um tipo de SGBD específico para o gerenciamento de triplas RDF, denominados de *RDF Store*, *Quadstore* ou pela forma mais comum, *Triplestore*.

O Triplestore adiciona suporte para o SPARQL e GeoSPARQL (no caso dos *Triplestore* espaciais). Uma das principais funcionalidades dos *Triplestores* é o suporte ao raciocínio ontológico (inferência) obtido nos resultados das consultas SPARQL, embora muitas implementações de *Triplestores* não dão suporte a essa funcionalidade.

A SPARQL possibilita a consulta de dados publicados em *Triplestores* por intermédio de interfaces conhecidas como *endpoints* SPARQL. São serviços Web que implementam o protocolo SPARQL para realização de consultas.

Existem diversos SPARQL *endpoints* públicos na WD, o mais conhecidos é a DBpedia⁴, que trata-se de um projeto que captura dados contidos nos *infoboxes*⁵ das páginas da Wikipédia⁶ e os mapeia para o modelo RDF com o uso de termos de diversas ontologias, como a ontologia da DBpedia⁷.

O W3C mantém uma página onde cita implementações de *Triplestores* de grandes capacidades, com previsões das capacidades de escalabilidade dessas implementações⁸.

⁴ <http://dbpedia.org/sparql>

⁵ são tabelas de formato fixo colocadas no canto superior direito dos artigos da Wikipédia e visa apresentar um resumo com aspectos relevantes e facilitar a consulta de outros artigos relacionados

⁶ <https://www.wikipedia.org/>

⁷ <http://dbpedia.org/ontology/>

⁸ <http://www.w3.org/wiki/LargeTripleStores>

Contudo, para a implementação da nossa infraestrutura semântica, iremos adotar tecnologias de *Triplestores* para armazenamento das triplas resultante dos mapeamento dos dados sobre biodiversidade com termos de ontologias de biodiversidade. Por meio de consultas, no padrão SPARQL e GeoSPARQL, os usuários poderão explorar, além de seus dados mapeados, dados provenientes do raciocínio ontológico (*reasoning*) e geoespacial.

3.3 *Linked Data – LD*

Segundo (HEATH; BIZER, 2011), a Web atual deixou de ser apenas um espaço global de documentos interligados entre si e está se tornando uma WD enorme com dados vinculados, constituído de bilhões de triplas RDF que cobrem os mais variados domínios do conhecimento humano.

A idéia do LD, ou dados vinculados, é mais recente do que a Web Semântica. O LD não é uma especificação ou recomendação oficial, trata-se de um conjunto de melhores práticas para a construção de uma infraestrutura de dados para facilitar o compartilhamento de informação em toda a WD.

O LD utiliza-se das tecnologias da Web Semântica, como RFD-S, OWL e SPARQL, para criar aplicativos consumidores de dados (DUCHARME, 2011). O termo LD refere-se a um conjunto de melhores práticas para a publicação e interligação de dados estruturados na WD. Essas práticas foram introduzidas por Tim Berners-Lee e tornaram-se conhecidas como os princípios do LD (BERNERS-LEE, 2006). Esses princípios são os seguintes:

- **Usar URI como nomes para coisas** – usar uma URI é a melhor maneira para identificar coisas e, portanto, é também a melhor maneira para identificar conexões entre elas.
- **Usar URI HTTP para que as pessoas possam procurar esses nomes** – usar prefixos de outros protocolos diferentes do HTTP para URI, como por exemplo *ftp:* e *ssh:*, diminuem a interoperabilidade, impossibilitando atingir os objetivos do *Linked Data*.
- **Quando alguém procurar por uma URI, prover informação útil, usando os padrões (RDF, SPARQL)** – uma URI pode ser apenas um nome e não realmente um endereço para uma página Web válido, mas aplicando este principio pode-se colocar informações relevantes nessas URI, utilizando padrões reconhecidos pela *Linked Data* como RDF, RFD-S, OWL e SPARQL.
- **Incluir links para outras URI, de modo que possam permitir a descoberta de mais coisas** – com base no principio anterior, existem vários vocabulários em RDF que fornecem outras propriedades para dizer que o recurso representado por uma determinada URI, ou esse elemento de dados, tem uma relação específica com outro recurso na Web. Dessa maneira, pode-se navegar por estes *links* e encontrar novas informação relevantes.

Fonte: Adaptada de [Cyganiak e Jentzsch \(2011\)](#).

Figura 21 – O diagrama da nuvem de dados presentes na LOD

Esses princípios foram amadurecidos e passaram a ser adotados no desenvolvimento das comunidades de LD. Inúmeras iniciativas voltadas para fomentar a criação da WD surgiram nos últimos anos, uma delas é o projeto LOD do *Semantic Web Education and Outreach Interest Group* (SweoIG) da W3C ([W3C, 2013c](#)).

O objetivo do LOD é estender a Web com a publicação de conjuntos de dados abertos na Web, por meio do modelo RDF, e estabelecendo ligações (relações) entre as entidades presentes em diferentes conjuntos de dados. As ligações realizadas com o modelo RDF permite relacionar entidades presentes em um conjunto de dado distinto com outras entidades de conjuntos de dados da LOD. Estas ligações poderão ser seguidas por agentes de software, como navegadores da Web Semântica e por *crawlers*⁹ de máquinas de busca, e assim aumentar o alcance das informações conhecidas de um determinado recurso na Web, além de fornecer um sofisticado sistema de busca, descoberta e rastreamento de informações relevantes para consultas distribuídas.

Dessa maneira, os dados abertos e interligados, usando o modelo RDF, poderão ser consumidos, de forma aberta, por diversas aplicações na Web, acessando os arquivos RDF e consultando os *endpoint* SPARQL.

A Figura 21 apresenta os conjuntos de dados e suas respectivas interligações que estão presentes na LOD, publicados por diversos projetos até setembro de 2011. Havia cerca de 295 registros de conjuntos de dados, com mais de 31 bilhões de triplas, com cerca de 504 milhões de interligações (reuso de URI e mapeamentos de URI para o mesmo recurso) ([CYGANIAK;](#)

⁹ programa de computador que navega pela rede mundial de uma forma metódica e automatizada

JENTZSCH, 2011).

A infraestrutura semântica proposta pretende seguir os conceitos do LD, assim poder disseminar os dados sobre biodiversidade por meio da LOD. Assim os usuários de programas e projetos de pesquisa, por exemplo, poderão combinar seus dados com outros conjuntos de dados presentes na LOD e dessa maneira poderão realizar consultas com resultados mais abrangentes.

3.4 Mapeando arquivos tabulados para o modelo RDF

A Web Semântica fornece tecnologias que permitem à integração e a disseminação de dados científicos.

Na área de IB, apresentada na seção 2, grande parte dos dados sobre biodiversidade estão disponibilizados em formatos de arquivos tabulados, como planilhas, CSV, TSV, etc. Para utilizar as tecnologias da Web Semântica é necessário mapear estes dados para o modelo RDF. Além disso, seguir os princípios do LD permitirá o aumento da abrangência das informações presentes nas entidades biológicas, descritas nos dados sobre biodiversidade, com os dados presentes na WD e na iniciativa LOD (BERNERS-LEE, 2006; SHADBOLT; BERNERS-LEE; HALL, 2006).

Em geral, existem duas abordagens para mapear os dados tabulados para o modelo RDF, o mapeamento baseado em linhas e o baseado em células.

O mapeamento baseado em linhas é a abordagem mais comum. Nela é assumido que cada linha descreve um Recurso (ou o Sujeito da tripla) (S) e que cada coluna representa uma Propriedade (P), onde os valores das células (do arquivo tabulado) são o Objeto (O), formando uma tripla (S,P,O) (DIMOU *et al.*, 2014). Esta abordagem permite obter rapidamente um modelo RDF a partir de qualquer arquivo tabulado.

O mapeamento baseada em células pressupõe que cada linha poderá representar mais de um Recurso/Sujeito (S), que as Propriedades (P) poderão originar-se de vocabulários e/ou ontologias (externos ao arquivo tabulado) e que o Objeto (O) poderá, também, originar-se de vocabulários e/ou ontologias, ou apontar para outros valores distintos de Recurso/Sujeito (S) (presentes na linha), ou corresponder-se a valores literais presentes nas células do arquivo tabulado, com tipos de dados predefinidos (datas, números, texto, etc.). Assim, formarão as triplas (S,P,O) resultante deste tipo de mapeamento.

Atualmente, a W3C disponibiliza uma lista¹⁰ de ferramentas para mapeamento de dados tabulados para o modelo RDF.

Foi realizado uma revisão na literatura com base em trabalhos relacionados ao mapeamento de dados tabulados para o modelo RDF que estivessem estreitamente relacionados a proposta deste trabalho, uma vez que a maior parte dos dados sobre biodiversidade são disseminados em formatos de arquivos tabulados.

¹⁰ <http://www.w3.org/wiki/ConverterToRdf>

Foram selecionadas algumas ferramentas relevantes a proposta deste trabalho, como as recomendações do RDB2RDF¹¹, R2RML, o *Apache Any23*¹², o *RDF Refine*¹³, o *SemantEco Annotator*, o *csv2rdf4lod*¹⁴, a *SML*¹⁵.

Uma visão mais geral das ferramentas de mapeamento para fontes estruturadas de dados, como CSV, é apresentado em (UNBEHAUEN *et al.*, 2012).

O grupo de trabalho RDB2RDF¹⁶ do W3C publicou duas recomendações para o mapeamento de banco de dados relacionais e arquivos de dados tabulados para o modelo RDF (SCHARFFE; ATEMEZING; TRONCY, 2012). A primeira recomendação é o mapeamento direto dos dados para RDF¹⁷, que é semelhante à tradução baseada em linha. Esta abordagem permite obter rapidamente um modelo RDF a partir de qualquer arquivo tabulado, desde que o arquivo siga a norma RFC4180¹⁸, em que o arquivo deverá fornecer um cabeçalho para nomear as suas colunas, ou seja, os valores contidos na primeira linha do arquivo de texto tabulado serão considerados o nome das colunas (que serão as propriedades da tripla). A segunda recomendação utiliza a linguagem R2RML¹⁹ para realizar o mapeamento dos arquivos e tabelas de banco de dados para o modelo RDF, além de permitir o reuso de termos provenientes de vocabulários controlado e ontologias externas aos dados. Os mapeamentos feitos na linguagem R2RML são descritos, propriamente, como um grafo RDF, utilizando a sintaxe *Turtle* do RDF (STADLER *et al.*, 2015).

O *Apache Any23* (ou do inglês *Anything To Triples*) trata-se de um conjunto de ferramentas desenvolvidas para mapear uma variedade de tipos de documentos na Web (HTML, Microformatos, CSV, etc.) para o modelo RDF. Esta ferramenta é composta por uma biblioteca, um serviço Web e uma ferramenta de linha de comando e é usado em um grande número de aplicações. O *Apache Any23* mapeia, também, arquivos tabulados, como o CSV, para o modelo RDF, seguindo um algoritmo de extração²⁰. Este algoritmo requer que:

- Os arquivos CSV sejam compatíveis com o padrão RFC4180, os cabeçalhos das colunas serão utilizado como propriedades da tripla RDF;
- Uma URI base deve identificar o arquivo CSV.

No entanto, para executar a interconexão dos registros de arquivos CSV com outras entidades na WD, o usuário deve colocar URIs de suas entidades em uma nova coluna no arquivo CSV

¹¹ <http://www.w3.org/2001/sw/rdb2rdf>

¹² <http://any23.apache.org/>

¹³ <http://refine.deri.ie>

¹⁴ <https://github.com/Timrdf/csv2rdf4lod-automation/wiki>

¹⁵ <http://sparqlify.org/wiki/SML>

¹⁶ <https://www.w3.org/2001/sw/rdb2rdf/>

¹⁷ <http://www.w3.org/TR/rdb-direct-mapping/>

¹⁸ <http://www.ietf.org/rfc/rfc4180.txt>

¹⁹ <http://www.w3.org/TR/r2rml/>

²⁰ <http://any23.apache.org/dev-csv-extractor.html>

original.

O *RDF Refine* é uma extensão do *OpenRefine*²¹, onde os dados podem ser convertidos em RDF (MAALI; CYGANIAK; PERISTERAS, 2011). O *RDF Refine* exige que os usuários definam um esqueleto, por meio de sua interface gráfica, para mapear os dados para o modelo RDF, como uma estrutura de árvore. Dessa forma, a ferramenta cria o mapeamento com os dados das colunas do arquivo com os nós da árvore (esqueleto). O principal recurso do *RDF Refine* é a ferramenta de reconciliação de dados. Esta ferramenta permite que o usuário execute buscas para reusar URI de entidades conhecidas na WD, como a DBpedia²², dessa forma, associar estas URI as entidades correspondentes do arquivo tabulado (MAALI; CYGANIAK; PERISTERAS, 2011). No entanto, este procedimento requer alguns ajustes, como a transposições e adições de novas colunas, alterando assim a estrutura do arquivo tabulado original.

O *SemantEco Annotator* é semelhante ao *RDF Refine*, pois oferece uma simples interface gráfica com o usuário, mantendo todas as interações para mapear o RDF separando do conteúdo da tabela de dados original, que permanece inalterada (SEYED *et al.*, 2013). Esta ferramenta serve de *front-end* de outra ferramenta de mapeamento, o *csv2rdf4lod*.

O *csv2rdf4lod tool* (LEBO; WILLIAMS, 2010) converte arquivos de dados tabulares, como CSV, para RDF de acordo com a interpretação de parâmetros codificados por meio de um vocabulário específico para conversão²³. Esta ferramenta tem dois tipos de mapeamento: bruto e aprimorado. O mapeamento bruto utiliza a tradução baseada em linhas, enquanto o mapeamento aprimorado precisa de um arquivo RDF codificado em um vocabulário específico para fazer a conversão. Este arquivo RDF também fornece o histórico de provas das operações que os dados originais sofreram ao se tornarem um modelo RDF.

A *Sparqlification Mapping Language* (SML) é uma linguagem de mapeamento do tipo RDB2RDF, com expressividade compatível com a R2RML, mas com uma abordagem sintática diferente (STADLER *et al.*, 2015). A SML possui uma estrutura de mapeamento legível para humanos. Esta linguagem cria modelos RDF virtuais a partir do mapeamento de bancos de dados relacionais ou de arquivos de texto tabulados. Esta linguagem faz parte da plataforma *Sparqlify*, que integra vários componentes de um aplicativo Web a um *endpoint* SPARQL.

A Tabela 1 apresenta uma comparação dos requisitos da infraestrutura proposta (discutidos na Seção 4.4) com as funcionalidades implementadas nas ferramentas de mapeamentos apresentados nesta Seção. Dentre as ferramentas citadas, apenas o *RDF Refine* faz a correspondência de instâncias para encontrar URIs de entidades na nuvem da WD e reutiliza-las no modelo RDF resultante do mapeamento, requisito reusar URIs da LOD.

Reutilizar URIs de entidades conhecidas em outros conjuntos de dados da WD é a chave principal para alcançar uma interoperabilidade semântica para o suporte de estudos

²¹ <http://openrefine.org>

²² <http://dbpedia.org>

²³ <http://purl.org/twc/vocab/conversion>

Tabela 1 – Comparação dos recursos entre as ferramentas de mapeamento CSV para RDF

Requisitos da Infraestrutura	RDF-R	SEA	csv2rdf4lod	AA23	R2RML	SML
Não modificar o arquivo original CSV	–	X	X	X	X	X
Verificar os tipos de dados do arquivo CSV	X	O	–	–	–	–
Adicionar Ontologias	–	O	X	–	X	X
Mostrar erros no mapeamento	–	O	–	–	–	–
Possuir editor gráfico	X	X	–	–	–	O
Mapeamento baseado em linhas	X	X	X	X	X	X
Mapeamento baseado em células	X	X	X	–	X	X
Mapeamento baseado em código	–	–	X	–	X	X
Permitir reuso dos mapeamentos	–	–	X	–	X	X
Reusar URI da LOD	X	–	–	–	–	–
Criar modelo RDF	X	X	X	X	X	X
Criar modelo OWL	–	–	–	–	–	–
Materializar as inferências	–	–	–	–	–	–
Exportar para <i>endpoint</i> SPARQL	–	–	–	–	–	X
Legenda: X Desenvolvido; O Em desenvolvimento; – Sem Informação.						

transdisciplinares e descoberta de novos conhecimentos. Um pesquisador poderá verificar os seus dados sobre um determinado recurso com os dados presentes na WD do mesmo recurso, enriquecendo o seu dado com novos fatos, além de poder atualizar suas informações a partir de outras fontes com maior credibilidade.

Além disso, nenhuma destas ferramentas afirma permitir que usuários criem um modelo RDF integrado com o OWL das ontologias (utilizadas no processo de mapeamento), como no requisito criar modelo OWL. O OWL resultante do processo de mapeamento poderia ser utilizado em ferramentas de desenvolvimento de ontologias, como o *Protégé*²⁴, ou por máquinas de inferências (*reasoners*) para verificar inconsistências semânticas entre os elementos da Ontologia com as novas triplas resultantes do mapeamento, como no requisito materializar as inferências.

Um outro recurso muito útil, mas inexistente na maior parte das ferramentas citadas acima, é o uso de *endpoint* SPARQL para armazenamento e consultas do modelo RDF resultante do mapeamento, como no requisito exportar para *endpoint* SPARQL. Este recurso permite, também, que outros usuários façam consultas nos conjuntos de dados, usando os padrões da Web Semântica como a linguagem SPARQL.

Outra característica desejável é a detecção de mapeamentos inconsistentes, seja por inconsistências ontológicas, detectadas no processo de *reasoning*, ou por erros de sintaxe da linguagem de mapeamento, quando utiliza-se linguagens baseadas em *scripts*. Esta característica favorecerá usuários não familiarizados com as tecnologias da Web Semântica, como os biólogos, onde os erros podem ser tratados pelos próprios usuários, melhorando os mapeamento e até os

²⁴ <http://protege.stanford.edu/>

dados, como erros de tipo de dados detectados durante o mapeamento dos dados.

Para abordar as lacunas remanescentes no processo de mapeamento de dados tabulados, apresentados na Tabela 1, especialmente dados sobre biodiversidade, foi desenvolvida uma linguagem denominada BioDSL, apresentada no Capítulo 4.

3.5 A Ontologia sobre Biodiversidade

A área de biodiversidade é um campo transdisciplinar de pesquisas, que requer a cooperação de pesquisadores de várias áreas do conhecimento humano. Biólogos podem realizar diferentes tipos de atividades, incluindo coleta de dados em campo, análise e síntese dos dados sobre os espécimes coletados, seus habitats e da correlação com outros seres vivos. Estas atividades geram dados, que vem sendo coletados em diversos lugares ao redor do mundo e publicados em formatos heterogêneos de dados, que podem não possuir padrão algum.

Não obstante, a semântica dos dados e informações presentes nos dados sobre biodiversidade não são capturadas de forma explícita. Ou seja, boa parte da interpretação do conhecimento contido nos dados é realizada pelos especialistas em biodiversidade.

Além do mais, dados sobre biodiversidade são heterogêneos e encontram-se dispersos em diferentes fontes, algumas de fácil acesso na Web, como artigos científicos, livros, relatórios técnico-científicos, dissertações e teses, e outras são de difícil acesso e localização, como diários de campo, planilhas, arquivos, pastas, etc.

A semântica no contexto de dados sobre biodiversidade é um fato que pode conter mais de uma interpretação, dependendo da visão e concepção dos pesquisadores. Por exemplo, tipo de vegetação também pode ser considerado um bioma.

Uma estratégia para tratar estes tipos de problemas é emprego de ontologias, que explicitamente definem os termos do vocabulário a ser utilizado, além de explicitar, também, os relacionamentos e hierarquias entres esses termos.

O LIS tem se dedicado ao desenvolvimento de uma ontologia para o domínio de biodiversidade, denominada de Ontologia de Biodiversidade (OntoBio) (ALBUQUERQUE, 2016; ALBUQUERQUE; SANTOS; de Castro Júnior, 2016; ALBUQUERQUE, 2011; SANTOS *et al.*, 2011). A OntoBio é baseada nos princípios da engenharia de ontologias e foi fundamentalmente baseada pela pesquisa de (GUIZZARDI, 2005).

O principal objetivo da OntoBio foi prover uma conceitualização clara e precisa dos aspectos relacionados as coletas de dados sobre biodiversidade, independente de uma aplicação específica (ALBUQUERQUE, 2011).

A OntoBio é composta por cinco sub-ontologias, como apresentado na Figura 22, conectadas por relações entre os conceitos e pelos seus axiomas.

Fonte: Adaptada de [Albuquerque \(2011\)](#).

Figura 22 – Visão geral da OntoBio

Sub-ontologia de Coleta – captura a estrutura de um protocolo de coleta, onde uma coleta deve necessariamente estar associada a um local de coleta, a uma instituição e a um objeto.

Sub-ontologia Entidade Material – conceitualiza as entidades materiais que fazem parte de uma coleta. Nesta sub-ontologia é composta por Entidade Abiótica e Entidade Biótica. A Entidade Abiótica caracteriza todos os fatos físicos que não possuem vida. A Entidade Biótica representa todos os fatores físicos vivos.

Sub-ontologia Localização Espacial – conceitualiza o contexto de espaços geográficos e de coordenadas geográficas, que tem relação com as coletas. Um espaço geográfico pode estar relacionado as várias coordenadas geográficas de localidades.

Sub-ontologia Ecosistema – conceitualiza os ecossistemas de acordo com os protocolos de coletas adotados. Um ecossistema é uma unidade natural com todos os elementos abióticos (“inanimados”) e bióticos (“animados”) e seus relacionamentos.

Sub-ontologia Ambiente – conceitualiza os ambientes que estão relacionados as coletas. Na biologia ambiente está relacionado a tudo que afeta diretamente o metabolismo e comportamento de um ser vivo, como luz, água, clima, etc.

Detalhes sobre o desenvolvimento e implementação da OntoBio podem ser encontrados em ([ALBUQUERQUE, 2011](#); [ALBUQUERQUE, 2016](#)).

De modo geral, uma ontologia estrutura o processo de aquisição de conhecimento para o propósito de compreensão, entendimento de máquinas e de extração de conhecimento do ambiente da Web Semântica. Esta tecnologia, aplicada ao domínio da biodiversidade, é um

recurso valioso para o planejamento estratégico do futuro da região Amazônica e sua contribuição para o planeta. Este trabalho utiliza a OntoBio, assim como outras ontologias relacionadas a área de biodiversidade para o mapeamento dos dados, onde os usuários poderão criar suas próprias bases de conhecimento.

3.6 Considerações Finais

As tecnologias da Web Semântica tornam dados passíveis de interpretação humana e por máquinas. Ademais, estas tecnologias oferecem a oportunidade de adicionar conteúdo semântico aos dados *on-line*, em especial aos dados sobre biodiversidade, permitindo que computadores possam fazer inferências a cerca dos mesmos, integrá-los e gerar novos conhecimentos (SANTOS *et al.*, 2011).

Tecnologias base da Web Semântica, como URI, RDF, RDF-S, OWL, SPARQL e LD, atualmente, possuem grande amadurecimento e uma grande comunidade de utilizadores e desenvolvedores. Essas tecnologias vem revolucionando diversas áreas do conhecimento, principalmente nas ciências da vida. No entanto, áreas como a IB pouco tem explorado estas tecnologias para integração e disseminação de dados.

Este trabalho aborda algumas lacunas para integração e disseminação de dados sobre biodiversidade, amenizando os esforços que os usuários destes dados têm ao utilizar as tecnologias da Web Semântica. Acredita-se que um ambiente amigável e integrado, em um alto nível de abstração, possa fazer com que usuários de dados sobre biodiversidade criem suas próprias bases de conhecimento, mapeando seus dados e integrando com outros dados e com a LOD, ampliando, assim, a abrangência destes dados e geração de novos conhecimentos, além de poder aumentar a participação de dados sobre biodiversidade na LOD.

Para tanto, é necessário criar formas para o mapeamento dos dados sobre biodiversidade para o modelo RDF. Apesar da literatura e da comunidade da Web Semântica possuírem ferramentas para mapeamento de dados, muitas lacunas ainda permanecem, principalmente para usuários com baixo domínio de computação, como os biólogos. Além do mais, as ferramentas atuais não seguem os princípios de LD durante o processo de mapeamento, uma atividade nada trivial em conjuntos de dados extensos, como os dados de biodiversidade. No Capítulo 4 é apresentada uma abordagem para o mapeamento de dados tabulados para o modelo RDF, a BioDSL, desenvolvida como uma linguagem de domínio específico para ser utilizada por biólogos e curadores de dados sobre biodiversidade.

BIODSL: UMA DSL PARA MAPEAMENTO DE DADOS SOBRE BIODIVERSIDADE

4.1 Considerações Iniciais

Nos últimos anos, dados abertos sobre biodiversidade tornaram-se amplamente disponíveis *on-line*, em repositórios de dados na Web, como GBIF¹ e *speciesLink*². Embora os repositórios de dados *on-line* tenham ajudado a aumentar a quantidade de dados sobre biodiversidade disponíveis na Web, muitas vezes, o significado agregado aos dados é ambíguo (MOURA *et al.*, 2012). Além do mais, esses dados estão armazenados em arquivos de texto (como o formato CSV), que não possuem informações semânticas explícitas (WAAL *et al.*, 2014).

Para enfrentar problemas correlatos, um movimento recente, denominado de LD, vem utilizando de tecnologias da Web Semântica para vincular e publicar dados, via Web, em formatos compreensíveis por humanos e máquinas, como o modelo RDF. O LD refere-se a um conjunto de melhores práticas para a publicação de dados na Web. Seguindo os princípios do LD, dados são associados a informações semânticas, integrados a outros conjuntos de dados distintos, pelo reuso de identificadores únicos – URIs, e processáveis por máquinas (BERNERS-LEE, 2006).

O LD usa o modelo RDF para representar o conhecimento em uma estrutura simples. No modelo RDF os dados são representados como pequenas declarações. Cada declaração é representada por um tripla (S, P, O), onde “S” é um URI para um determinado recurso na Web (podendo ser a representação de qualquer coisa conhecida, como uma pessoa, nação, espécie, etc.), o “P” é URI para determinar o predicado da tripla (normalmente é originário de um vocabulário ou ontologia, que contém a semântica explícita) e “O” pode ser um URI (um outro recurso no qual o “S” da tripla se relaciona de acordo com a semântica contida no respectivo “P” da tripla) ou um valor literal (ou qualquer dado associado a algum tipo de dado, como numeral,

¹ <http://www.gbif.org/>

² <http://splink.cria.org.br/>

texto, data, etc.) para expressar o valor a ser atribuído ao “S” da tripla.

Atualmente, o *Linked Open Data Project*³ (LOD) da W3C é o maior projeto que emprega LD para publicar dados na Web, resultando em uma nuvem de dados, denominada de WD. A WD contém uma grande quantidade de dados abertos, que abrangem uma grande variedade de disciplinas, incluindo as ciências da vida.

Embora exista uma grande quantidade de dados biológicos na WD, nenhuma quantidade significativa de dados sobre biodiversidade está disponível, para consulta, reutilização e compartilhamento.

Para publicar conjuntos de dados sobre a biodiversidade na WD, é necessário mapear estes dados para o modelo RDF, seguindo os princípios do LD. Ferramentas de mapeamento de arquivos tabulados para o modelo RDF (seção 3.4) oferecem pouco suporte para vincular os dados com a WD de forma automática ou semi-automática. Outras aparentes desvantagens encontradas nestas ferramentas incluem a falta de geração de modelos OWL integrado às ontologias utilizadas nos mapeamentos e com suporte ao raciocínio lógico (*reasoning*), à falta de suporte a publicação automática em *endpoint* SPARQL dos dados resultantes do mapeamento, falta de recursos para realizar a análise de qualidade e detecção de erros semânticos (durante o processo de *reasoning*) e de sintaxe nos mapeamento, além de poucas dessas ferramentas serem de fácil uso e configuração pelos usuários, como os biólogos que, em geral, são pouco familiarizados com as tecnologias da Web Semântica.

A natureza das informações sobre biodiversidade é dinâmica, muitas informações são de interpretação humana e podem variar durante o tempo, serem atualizadas ou caírem em desuso. Por estas características, para a construção da Infraestrutura Semântica propostas é necessário que os biólogos tenha uma ferramenta de mapeamento de dados simples e que possam realizar modificações em suas bases de conhecimento sem a necessidade de intervenção de especialistas em TI.

Este Capítulo apresenta a linguagem BioDSL, uma nova abordagem, baseada em uma *Domain-Specific Language* (DSL), para o mapeamento de dados sobre a biodiversidade (no formato CSV) para o modelo RDF. A BioDSL contempla as desvantagens das linguagens de mapeamento (seção 3.4), oferecendo aos usuários (biólogos em geral) um ambiente transparente, encobrindo parte da complexidade do processo de mapeamento baseado nos princípios da LD. A BioDSL é uma parte essencial deste trabalho, que visa a construção de uma infraestrutura semântica para dados sobre biodiversidade.

O restante deste Capítulo está estruturado em mais sete seções. A seção 4.2 apresenta um resumo sobre as linguagens de domínio específico. A seção 4.3 apresenta os recursos e funções da BioDSL. A seção 4.4 apresenta uma comparação da BioDSL com as outras linguagens de mapeamento. A seção 4.6 apresenta a lista de palavras reservadas e objetos da BioDSL. A seção

³ <https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

4.7 apresenta um exemplo de uso da BioDSL. A seção 4.8 apresenta a forma de utilização no ambiente terminal de comando BioDSL. A seção 4.9 apresenta conclusões sobre os trabalhos realizados no desenvolvimento e testes da BioDSL.

4.2 Linguagem de Domínio Específico

Em desenvolvimento de software, uma linguagem de domínio específico, do inglês *Domain-Specific Language* (DSL), é um tipo de linguagem de programação dedicada a um domínio específico de problema, usando expressões particulares ao problema e próprias dos especialistas do domínio, que também são seus usuários (ou programadores) (FREDDY; SU, 2017). Uma DSL fornece uma notação adaptada ao domínio de um aplicativo e é baseada nos conceitos e recursos relevantes desse domínio (Van Deursen; KLINT, 2002).

A ideia por trás das DSL é que um usuário, especialista de um domínio específico, pode programar nestas linguagens, sem qualquer experiência em desenvolvimento de software, instruindo o computador a realizar suas atividades. As DSL oferecem ganhos substanciais de expressividade e facilidade de uso em comparação com linguagens de programação de propósito geral, ou do inglês *General-Purpose Languages* (GPL), em seu domínio de aplicação (MERNIK; HEERING; SLOANE, 2005).

O conceito de DSL não é novo, existe desde o começo das linguagens de programação, mas o termo DSL tornou-se padrão devido à ascensão da modelagem de domínio específico (FOWLER, 2010). Para Dearle (2010), DSL é uma concepção para algo que os programadores profissionais utilizam todos os dias, existem muitos aplicativos que não podem ser totalmente escritos em uma única GLP. Os programadores são consumidores de diferentes DSL, cada qual com suas especificidades. Por exemplo, um programador programa a lógica do negócio na linguagem Java, usa a linguagem *Cascading Style Sheets* (CSS) para formatar a aparência de uma página Web, utiliza a linguagem *JavaScript* para entregar conteúdos dinâmicos a página Web e utiliza as linguagens Maven ou Gradle para criar *scripts* que configuram todo o ambiente do sistema e desenvolvimento.

As DSL podem ser classificadas de acordo com sua implementação, como Interna e Externa (FOWLER, 2010):

Interna: é o tipo de DSL que usa a infraestrutura de uma outra linguagem de programação existente (também conhecida como linguagem *host* da DSL) para criar uma semântica específica do domínio a ser suportado pela DSL. Na maioria dos casos, uma DSL interna é implementada como uma biblioteca sobre a linguagem *host* existente. Rails é a DSL interna mais conhecida e é implementada usando a infraestrutura da linguagem Ruby, ou seja, é uma DSL que foi construída para desenvolver aplicativos Web.

Externa: é o tipo de DSL desenvolvida com uma infraestrutura separada para: análise

léxica; técnicas de parse; interpretação; compilação e geração de código. Este tipo de DSL possui uma sintaxe independente de outras linguagens de programação. O desenvolvimento de uma DSL do tipo Externa é semelhante ao desenvolvimento de uma nova linguagem de programação a partir do zero, ou seja, com sintaxe e semântica próprias. No entanto, na maior parte dos casos, não será necessário implementar todas as complexidades de uma linguagem completa, apenas o que realmente for necessário para o domínio ao qual se destina a DSL.

A BioDLS é a linguagem desenvolvida para o domínio de mapeamento para o modelo RDF de dados sobre biodiversidade. A implementação da BioDSL foi do tipo Interna, pois utiliza a plataforma da linguagem *Groovy* como sua linguagem *host*. A linguagem *Groovy* possui recursos que dão suporte a implementação de DSLs e utilização de bibliotecas da linguagem Java. A plataforma Java possui diversas bibliotecas que dão suporte as tecnologias da Web Semântica, trazendo praticidade ao desenvolvimento da BioDSL.

4.2.1 Linguagem Groovy

Na linguagem de programação Java, assim como em qualquer GPL, não existem limitações que impeçam o desenvolvimento de qualquer tipo de solução. No entanto, o estilo de programação orientado a objeto de Java faz com que a linguagem demande muitas configurações e codificações adicionais. Além do mais, Java não permite um estilo de programação dinâmico, como oferecidos pelas linguagens de programação Python e Ruby (DEARLE, 2010; KOENIG *et al.*, 2015).

A linguagem *Groovy* oferece suporte a recursos de programação dinâmica para a plataforma Java (KOENIG *et al.*, 2015). *Groovy* consegue reutilizar as bibliotecas Java existentes, pois é executada diretamente na *Java Virtual Machine* (JVM). Além disso, *Groovy* tem uma sintaxe parecida com a linguagem Java e possui suporte para programação funcional, produz *bytecodes* que são executados pela JVM, possui suporte a execução e interpretação de *scripts* e é interoperável (KOENIG *et al.*, 2015).

Groovy tem suporte nativo ao desenvolvimento de DSLs, que se integra facilmente à própria linguagem, de modo que não é aparente onde o código em *Groovy* termina e a DSL começa (DEARLE, 2010). As DSLs em *Groovy* podem rodar diretamente na JVM e acessar bibliotecas Java.

O benefício que a plataforma *Groovy* ofereceu para o desenvolvimento da BioDSL foi o suporte a sua natureza dinâmica no desenvolvimento de DSL, além de poder interagir com bibliotecas em Java, onde já existem diversas bibliotecas com suporte às tecnologias da Web Semântica, como a OWL API⁴, Apache Jena⁵, RDF4J⁶, *triplestores*, entre outras. O suporte a bibliotecas Java permitiu que a BioDSL fosse capaz de manipular mais facilmente as tecnologias

⁴ <http://owlapi.sourceforge.net/>

⁵ <https://jena.apache.org/>

⁶ <http://rdf4j.org/>

da Web Semântica, que fazem parte do escopo deste trabalho, como RDF, OWL, *reasoning*, *triplestore*, etc. Dessa maneira, o acesso e manipulação destas tecnologias foram encobertas, utilizando apenas o vocabulário específico da BioDSL, linguagem desenvolvida para usuários especialistas no domínio da Biodiversidade e da Web Semântica.

Estas características permitiram que a BioDSL fosse implementada como uma DSL Interna, estendendo a linguagem *Groovy* e consequentemente a linguagem Java, mas mantendo uma sintaxe próxima à linguagem usada pelos especialistas dos domínios de Biodiversidade e Web Semântica. Dessa forma, estes especialistas podem se tornar, na prática, programadores de seus próprios conjuntos de dados.

4.3 BioDSL: uma DSL para mapear dados sobre biodiversidade para o modelo RDF

BioDSL é um acrônimo para linguagem de domínio específico para mapeamento de dados sobre biodiversidade para o modelo RDF. O mapeamento dos dados, na BioDSL, segue as recomendações de LD. O propósito da BioDSL é tornar transparente, para os usuários finais (biólogos), o complexo processo de mapeamento de dados tabulados em CSV para o modelo RDF, contemplando desde a verificação de tipo de dado ao raciocínio ontológico em OWL, além de publicar e disponibilizar o resultado do mapeamento por meio de um *endpoint* SPARQL.

Além disso, a BioDSL permite a interligação entre as entidades presentes nos dados, no formato CSV, com as mesmas entidades presentes em outros conjuntos de dados na WD. Dessa maneira, a BioDSL propicia a integração de dados, por meio do reuso de URI para identificar as entidades biológicas presentes em diferentes conjuntos de dados sobre biodiversidade. Assim, a BioDSL pode propiciar a descoberta de novos conhecimentos sobre as entidades conhecidas em outros conjuntos de dados sobre biodiversidade e áreas afins disponíveis na WD.

Atualmente, dados sobre biodiversidade são amplamente disseminados por grandes redes mundiais, como o GBIF, que utilizam o padrão DwC-A (seção 2.2.1). Este padrão é composto por um arquivo tabulado, no formato CSV, associado a um arquivo de metadado em EML, no formato XML. O DwC-A é um meio fácil e popular para o compartilhamento de dados sobre biodiversidade. Sua popularização foi impulsionada por sua adoção por novas ferramentas de publicação de dados, como IPT (seção 2.4.4) (WIECZOREK *et al.*, 2012).

Em um típico arquivo de texto no formato CSV do padrão DwC-A, cada coluna pode representar entidades de diferentes tipos (classes) e cada linha codifica várias entidades relacionadas entre si, como a espécie coletada, site da coleta, instituição, coleção biológica, etc. Essas entidades, também, podem aparecer repetidas em várias linhas do arquivo. Este formato não possui informações semânticas explícitas, sua interpretação e compreensão é realizada pelos especialistas de biodiversidade.

Tabela 2 – Comparação dos recursos da BioDSL com outras ferramentas de mapeamento CSV para RDF

Requisitos da Infraestrutura	RDF-R	SEA	csv2rdf4lod	AA23	R2RML	SML	BioDSL
Não modificar o arquivo original CSV	–	X	X	X	X	X	X
Verificar os tipos de dados do arquivo CSV	X	O	–	–	–	–	X
Adicionar Ontologias	–	O	X	–	X	X	X
Mostrar erros no mapeamento	–	O	–	–	–	–	X
Possuir editor gráfico	X	X	–	–	–	O	X
Mapeamento baseado em linhas	X	X	X	X	X	X	X
Mapeamento baseado em células	X	X	X	–	X	X	X
Mapeamento baseado em código	–	–	X	–	X	X	X
Permitir reuso dos mapeamentos	–	–	X	–	X	X	X
Reusar URI da LOD	X	–	–	–	–	–	X
Criar modelo RDF	X	X	X	X	X	X	X
Criar modelo OWL	–	–	–	–	–	–	X
Materializar as inferências	–	–	–	–	–	–	X
Exportar para <i>endpoint</i> SPARQL	–	–	–	–	–	X	X
Legenda: X Desenvolvido; O Em desenvolvimento; – Sem Informação.							

A BioDSL foi desenvolvida, no escopo da Web Semântica, para possibilitar o mapeamento e vinculação de dados científicos tabulados a informações semânticas e ontologias, utilizando o modelo RDF e seguindo os princípios do LD. Dessa forma, os dados poderão ser consultados e consumidos de forma transdisciplinar, pois poderão ser vinculados a outras entidades conhecidas na WD. Esta vinculação permitiu, também, a descoberta de novas informações relacionada as entidades presentes nos dados e na WD.

4.4 Comparação com outras ferramentas de mapeamento

A BioDSL contempla as desvantagens das linguagens de mapeamento apresentadas na seção 3.4, oferecendo aos usuários, biólogos em geral, uma sintaxe transparente, encobrindo parte da complexidade do processo de mapeamento baseado nos princípios da LD. A principal desvantagem, da maioria dessas soluções de mapeamento de dados, é a suposição de que cada linha de uma planilha descreve uma única entidade, como a tradução baseada em linha. Dados sobre biodiversidade, em geral, compilam mais de uma entidade por linha, inviabilizando a utilização destes tipos de linguagem de mapeamento. Além disso, as ferramentas existentes são exclusivamente focadas no mapeamento dos dados CSV para o modelo RDF e não tentam interconectar entidades dos arquivos CSV com entidades correspondentes na WD. Apenas o modelo RDF não é suficiente para os dados sobre biodiversidade, que necessitam de mecanismos automáticos ou semiautomáticos para sua vinculação a outros conjuntos de dados.

Uma comparação dos recursos presentes nas ferramentas de mapeamento de dados e dos requisitos para o funcionamento da infraestrutura é apresentada na Tabela 2. A lista de requisitos da infraestrutura incluem:

- **Não modificar o arquivo original** – A BioDSL não modifica o arquivo original CSV do usuário, como a Ferramenta RDF-R (seção 3.4). Toda configuração fica apenas no *script* de mapeamento. As URIs recuperadas da WD ficam armazenadas no RDF resultante do mapeamento, assim como toda a estrutura mapeada.
- **Verificar os tipos dos dados do arquivo** – A BioDSL, através da Função *defineType* (seção 4.5.1), realiza uma verificação de tipos de dados (numéricos, texto, data, etc.) e informa ao usuário quando um valor presente no seu arquivo CSV não é compatível com o tipo de dado indicado. Dentre as ferramentas, apenas o RDF-R realiza esse tipo de verificação.
- **Adicionar Ontologias** – A BioDSL possui a Função *addOntology* (seção 4.5.2) para carregar informações de arquivos de ontologias em OWL. O usuário poderá utilizar o vocabulário dessas ontologias para criar seus mapeamentos. Além disso, no final do mapeamento será exportado um arquivo em OWL contendo as ontologias juntamente com os dados mapeados. Esse arquivo poderá ser usado por máquinas de *reasoning*. Dessa maneira, a BioDSL pode detectar inconsistências semânticas presentes nos dados ou criadas por algum problema no mapeamento do usuário.
- **Mostrar erros no mapeamento** – A BioDSL informa quando um erro no mapeamento ocorre, informando a linha e o motivo do erro. Na Função *defineType* (seção 4.5.1), a BioDSL informa erros de inconsistência de tipos de dados no arquivo CSV, informando também a linha e coluna que o erro ocorreu.
- **Possuir editor gráfico** – Um editor gráfico é um recurso desejável em qualquer linguagem de programação. Para integrar a BioDSL à infraestrutura semântica, proposta nesta tese, foi desenvolvido um protótipo baseado na Web para que usuários especialistas em dados sobre biodiversidade e Web Semântica possam desenvolver seus *scripts* de mapeamento, testá-los e publicá-los em *endpoints* SPARQL (Capítulo 5).
- **Mapeamento baseado em linhas** – Assim como em todas as ferramentas de mapeamento, a BioDSL também possui o recurso de mapeamento baseado em linhas, onde cada linha irá representar uma entidade. Essa funcionalidade foi apresentada na seção 4.5.7 sobre a Função *map*.
- **Mapeamento baseado em células** – Das ferramentas apresentadas na seção 3.4, apenas a ferramenta AA23 não possui esta funcionalidade. A BioDSL permite que o usuário defina onde os dados, presentes em cada célula do arquivo CSV, irão ser utilizados, seja para formar uma parte da URI utilizada por uma entidade (Função *uri*, seção 4.5.5), para definir o valor de um Objeto (O) de uma tripla ou para criar uma string de busca semântica (Função *matching*, seção 4.5.6).
- **Mapeamento baseado em código** – A BioDSL é uma linguagem *script* baseada nos recursos de suporte ao desenvolvimento de DSLs da linguagem *Groovy* (seção 4.2.1). Os

scripts de mapeamento da BioDSL não necessitam ser compilados para *bytecodes* para serem executados. Eles são interpretados em tempo de execução pela linguagem *Groovy*. Toda as configurações de como as triplas deverão ser mapeadas ficam armazenadas nos *scripts* de mapeamento, sem gerar modificações desnecessárias no arquivo CSV.

- **Permitir o reuso dos mapeamento** – Um *script* de mapeamento pode ser mais facilmente compartilhado entre os usuários da BioDSL, pois os dados sobre biodiversidade seguem padrões como o DwC-A. Assim, um *script* de mapeamento da BioDSL pode ser facilmente estendido e reutilizado por outros usuários ou aplicados em outros conjuntos de dados.
- **Reusar URI da LOD** – Apenas o RDF-R implementa este tipo de recurso. No entanto, o RDF-R armazena as URI no arquivo CSV. A BioDSL não modifica o arquivo CSV, as URIs recuperadas pela busca semântica são mantidas apenas no RDF resultante do mapeamento.
- **Criar modelo RDF** – A BioDSL, assim como todas as outras ferramentas de mapeamento, exporta o RDF resultado do mapeamento para arquivos RDF. O usuário da BioDSL conta com diversos formatos para serialização do RDF, como *turtle*, RDF-XML, N3, etc.
- **Criar modelo OWL** – A BioDSL cria, no final do processo de mapeamento, um arquivo OWL com os modelos das ontologias carregadas pela Função *addOntology* (seção 4.5.2) juntamente com o RDF resultante dos mapeamentos das entidades do CSV. Esse arquivo pode ser utilizado em outras ferramentas, como Protégé⁷, e por máquinas de inferência da Web Semântica.
- **Materializar as inferências** – A BioDSL cria, também, um arquivo OWL (com os modelos das ontologias e do RDF resultante do mapeamento) utilizando máquinas de inferência e materializando as triplas inferidas. Inconsistências semânticas poderão ser detectadas durante o processo de *reasoning* e serão informadas aos usuários.
- **Exportar o modelo RDF para um endpoint SPARQL** – No final do processo de mapeamento, se o usuário definiu um endereço de *endpoint* SPARQL pela Função *setEndpoint*, então a BioDSL criará um repositório e carregará os RDF do mapeamento para o endereço do *endpoint*. Dessa maneira, outros *scripts* de mapeamento poderão usar a busca semântica nesse *endpoint* para criar interconectar esses dados (reuso de URIs de entidades conhecidas entre os conjuntos de dados).

4.5 Funcionamento da BioDSL

A BioDSL foi desenvolvida na linguagem de programação *Groovy*, sua linguagem *host*, e possui uma sintaxe declarativa baseada em objetos e declaração de funções com parâmetros e

⁷ <http://protege.stanford.edu/>

construções internas pré definidas.

Por exemplo, o arquivo CSV a ser mapeado pela BioDSL é representado por um objeto, denominado “*csv*”, cujo atributos referem-se aos nomes das colunas do arquivo, ou seja, quando o usuário for se referir a coluna “*institutionID*” deverá utilizar o atributo de mesmo nome do objeto “*csv*”, ficando como “*csv.institutionID*”.

A principal função da BioDSL é a “*Map*”, utilizada para criar o mapeamento para o modelo RDF com os dados do objeto “*csv*”, criando a estrutura de um tripla RDF.

As outras funções auxiliares definem como as URI para os Sujeitos da tripla serão construídas ou recuperadas de entidades conhecidas na WD. Por meio do reuso destas URI, no RDF resultante do mapeamento, é possível realizar a descoberta de novos fatos (tripas RDF) sobre as entidades presentes em outros conjuntos de dados na WD.

4.5.1 Configurando o arquivo CSV na BioDSL

Código-fonte 1 – BioDSL: Configurando o arquivo CSV

```
1: addCsv "/home/inpa/occurrenceINPA.csv"
2: ignoreRow 2
3: ignoreRow [1, 3, 5, 6]
4: defineType csv.catalogNumber, INTEGER
5: defineType csv.modified, CALENDAR, NOTNULL, template: "YYYY-mm
    -dd"
```

O código-fonte 1 apresenta a forma como um arquivo CSV deve ser configurado em um *script* de mapeamento da BioDSL. A linha 1 do código-fonte 1 apresenta o uso da Função *addCsv*, que deve receber como parâmetro uma *string* contendo o caminho para o arquivo na máquina do usuário, ou poderá informar uma URI contendo o caminho na Web do arquivo a ser baixado e mapeado pela BioDSL. A BioDSL é compatível apenas com arquivos que seguem o padrão RFC4180⁸, que requer que a primeira linha do arquivo contenha os nomes das colunas.

A Função *ignoreRow*, utilizada na linha 2 do código-fonte 1, serve para informar qual linha do arquivo CSV deverá ser ignorada durante o processo de mapeamento para o modelo RDF. Se o usuário preferir, poderá informar, também, uma lista de valores separados por vírgula, contendo os números correspondentes as linhas do arquivo CSV, que deverão ser ignoradas durante o mapeamento, como utilizado na linha 3 do código-fonte 1. Em *Groovy*, uma lista de valores é delimitada por colchetes e com os valores separados por vírgula, sintaxe que também foi adotada para definir listas de valores na BioDSL.

⁸ <http://www.ietf.org/rfc/rfc4180.txt>

A Função *defineType*, utilizada na linha 4 e 5 do código-fonte 1, serve para configurar o tipo de dado a ser verificado em uma determinada coluna do arquivo. Se o valor da coluna for mapeada para um Objeto da tripla, o RDF resultante levará o tipo de dado definido pela Função *defineType*, caso contrário o Objeto da tripla terá apenas o valor como um literal, que é o tipo padrão para os Objetos da tripla na BioDSL.

Na linha 4, a Função *defineType* recebe apenas os dois valores obrigatórios, o nome da coluna do arquivo CSV e o tipo de dados, no caso INTEGER, que é uma palavra reservada da BioDSL.

Já na linha 5, a Função *defineType* define o tipo de dado CALENDAR (palavra reservada da BioDSL, que se refere a tipos de dados para datas) para os valores contidos na coluna “*modified*”, que obrigatoriamente não deverá conter valores nulos (NOTNULL – palavra reservada da BioDSL para se referir a valores não nulos) e terá os valores formatados seguindo uma *string template*, onde “YYYY” são os valores para o ano com quatro dígitos, “mm” são valores do mês com dois dígitos e “dd” são valores para o dia com dois dígitos.

Caso um valor não siga o padrão de tipo de dado definido pela Função *defineType*, um erro fatal irá interromper o processo de mapeamento e uma mensagem será lançada e apresentada ao usuário informando o número da linha e o valor que resultou no erro, por exemplo, o parse para o tipo INTEGER encontrou um valor não numérico em uma determinada linha da coluna “*catalogNumber*”, então a BioDSL irá ser finalizada informando o valor, a linha e o tipo de dado que deveria ter o valor, assim o usuário poderá corrigir o valor errado no seu arquivo CSV e executar outra vez o seu *script* de mapeamento.

4.5.2 Carregando Ontologias e definindo Prefixos

Código-fonte 2 – BioDSL: carregando Ontologias e definindo prefixos

```
1: addOntology ontobio, "/home/inpa/ontobio.owl"  
2: addOntology dwcterms, "http://rs.tdwg.org/dwc/rdf/dwcterms.rdf"  
3: definePrefix bio, "http://ontobio.inpa.gov.br/ontology/"  
4: definePrefix dcterms, "http://purl.org/dc/terms/"
```

Um recurso de destaque presente na BioDSL é a possibilidade do usuário poder utilizar o vocabulário das ontologias para descrever o mapeamento dos seus dados para o modelo RDF. Para tanto, a BioDSL necessita carregar o arquivo OWL da ontologia e criar um objeto para o usuário poder referencia-lo em seu mapeamento.

A Função *addOntology* da BioDSL é encarregada de carregar o modelo (no formato OWL ou RDF) das ontologias. A forma como o usuário deverá carregar um arquivo de ontologia é apresentado no código-fonte 2. A Função *addOntology*, assim como a Função *addCsv*, necessitam do caminho para o arquivo ou de uma URI para poder carregar a ontologia para a BioDSL. O

primeiro parâmetro, como apresentado na linha 2 do código-fonte 2, é o nome do objeto que o usuário irá utilizar para acessar os termos da ontologia, neste caso “*ontobio*” será o objeto para a ontologia presente no caminho “*/home/inpa/ontobio.owl*”. Já a linha 3 do código-fonte 2 apresenta a segunda forma de localização do arquivo, ou seja, neste caso o objeto “*dwcterms*” irá carregar a ontologia presente no endereço de URI em “*http://rs.tdwg.org/dwc/rdf/dwcterms.rdf*”. Em ambos os casos, é necessário que o caminho passado no segundo parâmetro da Função, aponte para um arquivo OWL ou RDF referentes a ontologia.

Ontologias na Web Semântica, especialmente em OWL, poderão utilizar vários prefixos para simplificar seus termos. Os termos das ontologias são URI, alguns contêm a mesma raiz, então a raiz pode ser simplificada em um prefixo. O prefixo principal de uma ontologia é a sua URI base, onde a maior parte dos termos da ontologia se originam. Geralmente, uma ontologia em OWL possui vários prefixos diferentes. Para o usuário poder referenciar termos que não possuem o mesmo prefixo da URI base da ontologia, a BioDSL disponibiliza a Função *definePrefix*.

A Função *definePrefix* é utilizada nas linhas 3 e 4 do código-fonte 2. Esta Função é muito parecida com a Função *addOntology*. No entanto, só é admissível uma URI no segundo parâmetro. O primeiro parâmetro é o nome do objeto prefixo que guardará o valor da URI base informada no segundo parâmetro. Por exemplo, o objeto “*bio*” é um prefixo, se o usuário quiser acessar o termo “*Species*”, então ele deverá se referenciar a “*bio.Species*” (o mesmo que *http://ontobio.inpa.gov.br/ontology/Species*) em seu mapeamento.

4.5.3 Definindo uma URI base

Código-fonte 3 – BioDSL: Definindo uma URI base

```
1: uriBase "http://www.inpa.gov.br/resource/"
```

A BioDSL criará um modelo RDF ou OWL no final do processo de mapeamento, caso não ocorram erros citados acima. Este modelo deverá possuir sua URI base, que servirá como base a representação (uma URI) das entidades presentes nos dados CSV.

A Função *uriBase* permite aos usuários configurar um URI base para o modelo RDF a ser gerado.

Na linha 1 do código-fonte 3 é utilizada a Função *uriBase*. Esta Função requer apenas um parâmetro, que será uma URI para representar as entidades presentes nos dados do arquivo CSV. Por padrão, a BioDSL definirá a URI dos recursos (entidades) que serão mapeados para o modelo RDF utilizando-se desta URI base. Existem outras formas do usuário poder customizar um URI para uma(s) determinada(s) entidade(s), como é o caso das Funções *URL*, *Matching* e *Gazetteer*, apresentadas logo a seguir.

4.5.4 Configurando o Triplestore

Código-fonte 4 – BioDSL: Configurando o triplestore

```
1: useTripleStore "http://10.70.2.106:8080/sparql", BLAZEGRAPH
```

A Função *useTripleStore* serve para o usuário definir um endereço de *triplestore*. A BioDSL irá utilizar este endereço para carregar o RDF, resultante do processo de mapeamento. A Função *useTripleStore* requer dois parâmetros, a URI do *endpoint* SPARQL da *triplestore* e uma palavra reservada referente a implementação da *triplestore*. Atualmente a BioDSL suporta dois *triplestore* *Blazegraph*⁹ e *Virtuoso*¹⁰, com as respectivas palavras reservadas BLAZEGRAPH e VIRTUOSO.

Na linha 1 do código fonte 4 é apresentado um exemplo de uso desta Função. O primeiro parâmetro, "*http://10.70.2.106:8080/sparql*", indica o caminho para o *endpoint* SPARQL da *triplestore*. O segundo parâmetro BLAZEGRAPH informa qual é a implementação de *triplestore* que a BioDSL irá trabalhar. Atualmente existem diversas implementações de *triplestore*, com recursos distintos (como o *full text search*¹¹), que fazem com que os processos de manipulação e criação de um novo repositório pelo *triplestore* sejam realizados de formas diferentes em cada implementação de *triplestore*.

4.5.5 A Função URI

Código-fonte 5 – BioDSL: Função URI

```
1: uri occurrenceURI, "ocurrence", csv.occurrenceID
2: uri institutionURI, "institution", csv.institutionID
```

A BioDSL permite que o usuário customize uma URI para um determinado conjunto de entidades do arquivo CSV. A Função *uri* é utilizada para construir uma URI a partir da URI base da BioDSL. O usuário poderá utilizar os dados contidos nas colunas do arquivo CSV para criar URI customizadas.

A Função *uri* utilizada na linha 1 e 2 do código fonte 4 possui um parâmetro obrigatório, o nome do objeto *uri*, no caso *occurrenceURI* e *institutionURI*, e os demais parâmetros são opcionais. Este parâmetros opcionais podem ser strings, separadas por aspas, ou nome de colunas do arquivo. A Função *uri* concatenará à URI base os valores dos outros parâmetros, caso for um *string* ou o valor contido em uma determinada coluna do arquivo.

⁹ <https://www.blazegraph.com/>

¹⁰ <https://virtuoso.openlinksw.com/>

¹¹ <https://www.w3.org/2009/sparql/wiki/Feature:FullText>

Por exemplo, na linha 1 do código fonte 4 a URI resultante para o objeto *occurrenceURI* será "<http://www.inpa.gov.br/resource/>" (declarado na linha 1 do código fonte), mais a *string* "*ocurrence*", mais o valor contido na linha 1 do arquivo CSV na coluna *occurrenceID*, que poderia ser "0001", então, tudo isso resultaria na URI "<http://www.inpa.gov.br/resource/ocurrence/0001>".

4.5.6 A Função Matching

Código-fonte 6 – BioDSL: Função Matching

```
1: matching (institutionLOD) {  
2:     endpoint "http://triplestore.org/sparql"  
3:     vendor BLAZEGRAPH  
4:     search csv.occurrenceID  
5:     default_uri institutionURI  
6: }
```

A Função *matching* é um recurso importante da BioDSL. Esta Função permite vincular entidades presentes nos arquivos CSV com entidades correspondentes na WD, por meio de buscas semânticas realizadas em *endpoints* SPARQL da WD.

A linha 1, do código fonte 6, possui um exemplo de uso da Função *matching*.

Entre os parênteses fica o nome do objeto que armazenará os parâmetros da Função *matching* para realizar a busca semântica. Este objeto armazenará todos os parâmetros necessários para realizar a busca semântica em um determinado *endpoint* SPARQL na WD.

Entre as chaves ficam os nomes dos parâmetros e seus respectivos valores, cada par de parâmetro/valor deverá ser declarado em uma linha própria, não importando a ordem que o usuário deverá informa-los.

Para configurar os parâmetros da busca semântica, o usuário deverá declarar um conjunto de parâmetros com seus respectivos valores dentro dos limites das chaves da Função *matching*.

Alguns parâmetros são obrigatórios como:

- ***endpoint*** – deve ter um valor de *string* (entre aspas) com uma URI do *endpoint* SPARQL que será explorado pela busca semântica da BioDSL.
- ***vendor*** – atualmente a BioDSL é compatível apenas com duas implementações de *endpoint* SPARQL, *Blazegraph* e *Virtuoso*. Os valores aceitos por este parâmetro são palavras reservadas da BioDSL, respectivamente BLAZEGRAPH e VIRTUOSO.
- ***search*** – este parâmetro definirá a *string* de busca que será utilizada pelo mecanismo de busca semântica da BioDSL. Para criar uma *string* de busca, o usuário deverá informar o nome da coluna ou uma lista com nomes de colunas do arquivo CSV. A pesquisa será

realizada com base no conteúdo das células correspondentes aos nomes de colunas. Se uma lista com nome de colunas for passada como valor deste parâmetro, então o conteúdo recuperado de cada célula das colunas será concatenado, formando a *string* de busca.

Parâmetros opcionais da Função *matching*:

- **default_uri** – este parâmetro deverá receber um objeto da Função *uri* (Subseção 4.5.5) caso a busca semântica não retorne entidades, então uma URI poderá ser criada de acordo com o objeto *uri* associado a este parâmetro.
- **type** – este parâmetro serve para restringir a abrangência da busca, relacionando apenas entidades pertencentes a uma determinada classe. O usuário poderá informar o nome da classe utilizando um objeto da Função *addOntology* ou por um objeto da Função *definePrefix*, também, se preferir, o usuário poderá informar diretamente, por meio de uma *string* (entre aspas), uma URI correspondente a classe.

Como muitas entidades, presentes nos arquivos CSV, são conhecidas na WD, (*Geonames*, *DBpedia*, *SWI Gazetteer* (CARDOSO *et al.*, 2015), etc.) a BioDSL explora *endpoints* SPARQL realizando buscas semânticas sobre os Objetos literais (O) das triplas para encontrar valores correspondentes com as *strings* de busca. Para realizar a busca semântica, a BioDSL utiliza diferentes estratégias para diferentes implementações de *endpoint* SPARQL. Atualmente, a BioDSL suporta duas implementações de *endpoint* SPARQL: *Blazegraph*¹² e *Virtuoso*¹³.

O recurso de *Full Text Search* ainda não é um padrão oficial SPARQL, então diferentes *endpoint* SPARQL implementam formas distintas de realizá-lo. O *Full Text Search* retorna uma pontuação da similaridade entre a *string* de busca e os valores literais dos Objetos (O) das triplas RDF. Se o resultado da busca conter mais de uma entidade, então a entidade que tiver a melhor pontuação de *Full Text Search* será utilizada para criar o (S) da tripla. Se nenhuma entidade for retornada na pesquisa, então o objeto *uri* fornecido pelo parâmetro *default_uri* será utilizado para criar o (S) da tripla.

Por exemplo, se a *string* de busca for "*Harpia harpyja*" e o parâmetro *type* for a URI da classe "http://dbpedia.org/ontology/Species", então a consulta somente processará o *Full Text Search* para as entidades que pertencem a esta classe. A restrição de tipo de classe e a combinação da melhor pontuação de similaridade podem ser bastante eficazes para encontrar correspondências corretas, especialmente quando utiliza-se termos científicos, como nomes de espécies. No entanto, podem ocorrer erros nas correspondências, então os usuários deverão realizar testes na definição das strings de busca para obter aceitáveis taxas de erros de correspondências entre as entidades.

Com este mecanismo de busca semântica, a BioDSL é capaz de recuperar URIs de entidades correspondentes na WD para reutilizá-las no RDF resultante dos mapeamentos. Desta

¹² <https://www.blazegraph.com/>

¹³ <https://virtuoso.openlinksw.com/>

forma, a BioDSL afirma, em nível lógico, que as entidades da WD e do arquivo CSV (agora utilizando as mesmas URI) são equivalentes, assim uma vinculação é criada conectando a entidade do arquivo CSV à uma outra entidade na WD.

Essas correspondências podem ser facilmente enriquecidas, reutilizando e/ou incluindo os fatos já conhecidos (triplas RDF) sobre entidades da WD às triplas resultantes do mapeamento. Isso pode indicar um caminho para a descoberta de novos conhecimentos, explícitos e implícitos, nos dados da WD e dos arquivos CSV.

A reutilização de entidades dá ao RDF final um alto grau de interconectividade com a WD. Este recurso, também, permite criar interoperabilidade entre conjuntos de dados.

4.5.7 A Função Map

Código-fonte 7 – BioDSL: Função Map

```
1: map (linha) {
2:     rdfs.label csv.linha
3: }
4: map (institution, institutionLOD) {
5:     rdf.type ontobio.Institution
6:     rdfs.label csv.institutionCode
7: }
8: map (occurrence, occurrenceURI) {
9:     rdf.type dwc.Occurrence
10:    dwcterms.occurrenceID csv.occurrenceID
11:    dwcterms.catalogNumber csv.catalogNumber
12:    dwcterms.occurrenceRemarks csv.occurrenceRemarks
13:    dcterms.references csv.references
14:    ontobio.hasInstitution institution
15: }
```

Para cada tipo de entidade que o usuário deseja mapear de seu arquivo CSV, uma Função *map* deverá ser declarada. A Função *map* define como será a estrutura (S, P, O) das triplas RDF para as entidades do arquivo CSV. O (S) da tripla é definido pelo valor do parâmetro entre parenteses. Este parâmetro pode ser um objeto das Funções *matching* e *uri*, que são Funções desenvolvidas para determinar ou criar URIs para o (S) de triplas do mapeamento. Por padrão, caso o nome utilizado pelo usuário não corresponda a um destes objetos, a BioDSL irá realizar o mapeamento orientado por linhas, que determina que cada linha é uma entidade distinta, e a URI resultante para o (S) da tripla será a concatenação da URI base (seção 4.5.5), mais o nome declarado no parâmetro entre parenteses, mais o respectivo número da linha.

A linha 1 do código fonte 7, apresenta um mapeamento simples, onde o nome “linha” não faz correspondência a nenhum objeto previamente declarado (nos códigos fonte apresentados anteriormente). Por exemplo, a URI base é “*http://www.inpa.gov.br/resource/*”, declarada no código fonte 3, e o nome nos parenteses da Função *map* é “*linha*”, declarado na linha 1 do código fonte 7, então, a URI para entidade presente na primeira linha do arquivo CSV será “*http://www.inpa.gov.br/resource/linha:1*”. Por padrão, a BioDSL coloca o carácter dois pontos para separar o nome do parâmetro do respectivo número da linha do arquivo CSV.

Uma vez que uma URI para identificar uma determinada entidade é definida, ou seja o recurso (S) da tripla RDF, a BioDSL está pronta para definir suas respectivas Propriedades (P) a seus respectivos Objetos (O), formando assim triplas RDF (S, P, O).

As Propriedades (P) e os Objetos (O) da tripla (S) serão definidos nos limites das chaves da Função *map*. Cada linha, dentro dos limites das chaves, representa um par de Propriedade (P) e Objeto (O) da tripla a ser formada, separados por um espaço em branco.

Uma Propriedade (P) pode ser uma URI válida declarada como uma *string*, entre aspas, ou por um objeto da Função *addOntology* ou *definePrefix*. Na linha 10 do código fonte 7, para o usuário referenciar a propriedade “*occurrenceID*” da ontologia “*dwcterms*” deverá utilizar a seguinte sintaxe “*dwcterms.occurrenceID*”.

Para o Objeto (O) da tripla, o usuário deverá utilizar um nome de coluna do objeto *csv* ou um atributo do objeto da Função *addOntology*, ou *definePrefix*. Os Objetos (O) das triplas deverão ser precedido por suas respectivas Propriedade (P), assim como apresentado na linha 10 do código fonte 7 onde “*dwcterms.occurrenceID*” é a Propriedade (P) e o “*csv.occurrenceID*” é um nome de coluna.

O usuário poderá, também, utilizar um objeto das classes *uri*, *matching* e *map* para expressar um relacionamento entre entidades em seu mapeamento, onde o Objeto (O) será um URI, assim como na linha 14, a Propriedade (P) é “*ontobio.hasInstitution*” e o Objeto (O) é “*institution*”, que é um objeto de uma outra Função *map* (linha 4).

A BioDSL possui dois objetos padrões, o “*rdf*” e o “*rdfs*” que referenciam respectivamente os modelo RDF e RDF-S da W3C. Na linha 5 do código fonte 7 a Propriedade (P) “*rdf.type*” informa que a entidade “*institution*” pertence a classe “*ontobio.Institution*” da ontologia do objeto “*ontobio*” da Função *addOntology*, declarada na linha 1 do código fonte 2.

A BioDSL também permite que o usuário informe uma lista de objetos para um determinado predicado, a lista deverá ser delimitada por colchetes.

4.6 Palavras e objetos reservados

A BioDSL possui um conjunto de palavras e objetos reservados. O usuário da BioDSL não poderá utilizar estas palavras e objetos como nomes de seus objetos das Funções nos *scripts*

Tabela 3 – Palavras Reservadas da BioDSL

Palavra	Funções
BOOLEAN	<i>defineType</i>
CALENDAR	<i>defineType</i>
CHAR	<i>defineType</i>
DOUBLE	<i>defineType</i>
FLOAT	<i>defineType</i>
INTEGER	<i>defineType</i>
LONG	<i>defineType</i>
STRING	<i>defineType</i>
NOTNULL	<i>defineType</i>

Tabela 4 – Objetos Reservados da BioDSL

Objeto	Funções
BLAZEGRAPH	<i>useTripleStore</i> e <i>matching</i>
VIRTUOSO	<i>useTripleStore</i> e <i>matching</i>
csv	<i>defineType</i> , <i>matching</i> , <i>uri</i> e <i>map</i>
rdf	<i>map</i>
rdfs	<i>map</i>
a (<i>rdf.type</i>)	<i>map</i>

BioDSL de mapeamento. As Tabelas 3 e 4 apresentam, respectivamente, a lista de Palavras e Objetos reservados, bem como os nomes das Funções onde podem ser utilizados.

As palavras reservadas, Tabela 3, podem ser utilizadas no contexto da Função *defineType* (seção 4.5.1).

Os objetos reservados, Tabela 4, podem ser utilizados em diversos contextos das Funções, como o objeto *csv* que refere-se ao arquivo CSV, o *rdf* e *rdfs* que se referem a termos do RDF e RDF-S. O objeto *a* é um alias para o termo *type* do objeto *rdf*.

4.7 Um exemplo de uso da BioDSL

A Figura 23 apresenta um *script* BioDSL, um arquivo CVS, *endpoints* SPARQL e um arquivo RDF na sintaxe *turtle*. As setas vermelhas com rótulos numerados servem para demonstrar os respectivos recursos alvos ou resultados das Funções da BioDSL.

Para criar um mapeamento de dados CSV para o modelo RDF, o usuário deverá informar o caminho físico do seu arquivo CSV pela Função *addCSV* (seção 4.5.1). A linha 1 do *script* BioDSL possui um exemplo de uso da função *addCSV*, destacada pela seta de rótulo 1 apontada para o arquivo CSV, localizado ao lado direito superior da Figura 23.

Nas linhas 2 e 3 do *script* BioDSL é apresentado o uso da Função *ignoreRow* (seção 4.5.1), recebendo como parâmetro um valor numérico (linha 2) e uma lista de valores numéricos

(linha 3).

Na linha 4 do *script* BioDSL é demonstrado o uso da Função *defineType* (seção 4.5.1), que está ligada a coluna do arquivo CSV “*catalogNumber*” pela seta de rótulo 2. Caso um dos valores não possa ser parseado como valor numérico inteiro, a BioDSL apresentará um erro ao usuário com o número da linha, nome da coluna do arquivo CSV e o respectivo tipo de dado.

Na linha 6 do *script* BioDSL é demonstrado o uso da Função *uriBase* (seção 4.5.3) que recebe como parâmetro uma URI para o prefixo padrão do RDF resultante do mapeamento. Essa URI, também, será utilizada pela Função *uri* (seção 4.5.5) para criar as URIs das entidades dos arquivos CSV.

Na linha 7 do *script* BioDSL é demonstrado o uso da Função *useTripleStore* (seção 4.5.4). Esta Função define o endereço do *endpoint* SPARQL para a BioDSL armazenar o RDF resultante do mapeamento. Este *triplestore* poderá ser utilizado para integrar dados de outros usuários (por meio da busca semântica da BioSL – Função *matching*).

A Função *matching* (seção 4.5.6) é demonstrada na linha 14 do *script* BioDSL. O parâmetro *endpoint* deverá receber o endereço URI de um *endpoint* SPARQL. Na linha 15 o parâmetro *endpoint* recebeu um URI de um *endpoint* SPARQL (fictício) pertencente ao LOD, como é apresentado na ligação da seta de rótulo 5.

A principal Função da BioDSL é a *map*, pois é a única Função capaz de produzir RDF de fato, as demais Funções auxiliam esta Função a construir e recuperar URIs para os (S), carregar o vocabulário de termos das ontologias para serem utilizados na construção das Propriedades (P) e Objetos (O) das triplas RDF, definir os tipos de dados para os valores de (O), configurar o arquivo CSV, etc.

Um exemplo de uso da Função *map* é apresentado na linha 24 do *script* BioDSL. O (S) deste mapeamento é definido pelo objeto “*occurrenceURI*”, que é um objeto da Função *uri* da linha 12. A primeira Propriedade (P) do indivíduo é o termo *type* do objeto “*rdf*” (um objeto reservado da BioDSL para os termos do RDF).

O respectivo Objeto (O) desta tripla é a classe “*Occurrence*” do objeto “*dwcterms*”, declarado na linha 9 pela Função *addOntology* (seção 4.5.2).

Em resumo, a linha 25 do mapeamento “*occurrence*” afirma que o (S) “*occurrenceURI*” é um recurso pertencente a classe “*Occurrence*” (expressão dada pelo Predicado (P) “*rdf.type*”) da ontologia “*dwcterms*”, que forma o Objeto (O) da tripla.

Já a linha 26, o Predicado (P) é o termo “*occurrenceID*” da ontologia “*dwcterms*” e o Objeto (O) é um nome de coluna, “*occurrenceID*”, do arquivo CSV (objeto “*csv*”).

O usuário poderá criar relacionamentos entre os mapeamentos, como apresentado na linha 30. O Predicado (P) é o termo “*hasInstitution*”, da ontologia “*ontobio*”, e o Objeto (O) é um nome de mapeamento, “*institutuon*”, declarado na linha 20 do *script* BioDSL.

Fonte: Elaborada pelo autor.

Figura 23 – Configurando o arquivo CSV e mapeando-o para RDF em BioDSL

A seta de rótulo 6 liga a estrutura do mapeamento da Função *map*, delimitado por uma chave vermelha, aos seus respectivos RDF resultantes, na sintaxe *turtle* do RDF, também limitado a uma outra chave vermelha. Este RDF também será armazenado na *triplestore* setada na linha 7, como demonstrado pela ligação entre o arquivo RDF e o *triplestore* feita pela segunda seta de rótulo 3.

4.8 Executando a BioDSL no Terminal

A BioDSL necessita do ambiente de uma JVM instalada no computador do usuário (versão 1.8 ou superior) para poder realizar os mapeamento de arquivo CSV para o modelo RDF no ambiente de terminal. A JVM é multiplataforma, logo a BioDSL poderá ser executada pelos terminais dos Sistemas Operacionais GNU/Linux, Mac e Windows.

O usuário poderá acessar um arquivo CSV local ou remotamente, desde que o acesso ao link seja garantido pela rede a qual o seu computador esteja conectado.

Triplestores poderão ser utilizadas, desde que o acesso também seja garantido pela rede do usuário, para armazenamento do RDF resultante do mapeamento ou para realizar buscas semânticas da BioDSL e criar vinculações entre as entidades de conjuntos de dados.

O usuário poderá utilizar qualquer editor de texto, desde que salve os *script* BioDSL na codificação UTF-8. Com o acesso ao arquivo CSV garantido e juntamente com o arquivo de *script* BioDSL, o usuário poderá executar a BioDSL diretamente pelo terminal do GNU/Linux,

A linguagem possui uma sintaxe simples, com recursos que outras ferramentas de mapeamento não oferecem aos seus usuários, além de ter sido desenvolvida, a princípio, para usuários especialistas em biodiversidade.

A BioDSL proporciona aos usuários especialistas em biodiversidade mapear para o modelo RDF dados abertos sobre biodiversidade, como os do GBIF, *speciesLink* e SiBBR, que estão em formatos de planilha eletrônica e sem perdas de informações.

Os usuários poderão ser capazes de escrever, compartilhar e publicar seus *scripts* BioDSL juntamente com os arquivos CSV e triplas RDF (em arquivos no formato RDF e OWL ou por meio de *endpoint* SPARQL), dessa maneira, contribuirão para o crescimento da WD com dados sobre biodiversidade.

No contexto desta tese, a BioDSL é o coração da Infraestrutura Semântica proposta. Por meio da BioDSL, os usuários de biodiversidade poderão publicar, consultar, vincular e enriquecer seus dados com dados da WD ou com dados de outros usuários, em uma infraestrutura transparente. Elas esconde do usuário parte da complexidade envolvida no processo de mapeamento, vinculação e publicação de dados na WD.

No Capítulo 5 será apresentado como a BioDSL poderá ser utilizada para criar uma infraestrutura semântica para dados sobre Biodiversidade, por meio de um ambiente Web interativo, escondendo as configurações das funcionalidades e tornando esta linguagem mais amigável a usuários poucos experientes.

INFRAESTRUTURA SEMÂNTICA PARA DADOS SOBRE BIODIVERSIDADE

5.1 Considerações Iniciais

Nas ciências da vida, em geral, dados correspondem à uma coleção de fatos que exigem interpretação dos seus significados para então se tornarem conhecimento. Essa interpretação normalmente é realizada por humanos, com expertise na área. No entanto, para realizar a interpretação de grandes volumes de dados é imprescindível o uso de computadores, idealmente, com recursos de alto desempenho. Na área de biodiversidade, dados podem refletir fatos biológicos relacionados a outros fatos do conhecimento humano, que muitas vezes estão apenas implícitos no entendimento dos especialistas da área.

A IB é uma área de Informática aplicada ao desenvolvimento de sistemas de informação que proporcionem a interoperabilidade de dados e síntese do conhecimento em fontes locais de dados sobre biodiversidade, para então introduzi-las em arquiteturas de conhecimento globais com disseminação em grande escala na Web, como a rede do GBIF ([SHANMUGHAVAL, 2007](#)).

Já a área de Web Semântica atua na estruturação de dados e informações na Web, de forma que computadores possam compreender estes dados, fazer inferências lógicas a respeito dos mesmos, interligar os dados entre si e promover a interoperabilidade dos dados. O movimento do LD, ou dados vinculados, contribuiu para o surgimento de uma nova Web, a WD, composta por diversos dados vinculados entre si, provenientes de áreas distintas do conhecimento humano. Dessa forma, a WD oferece suporte à análise destes dados em uma abordagem transdisciplinar ([KAUPPINEN; ESPINDOLA; JONES, 2012](#)).

Neste contexto, as ontologias são empregadas para fornecer um vocabulário compartilhado, estruturar o conhecimento e permitir a interoperabilidade entre os conjuntos de dados na WD. Além disso, as ontologias formais facilitam a descoberta de conhecimentos implícitos, por

meio do processo de raciocínio lógico realizado pelas máquinas.

No entanto, apesar dos esforços da IB na criação de infraestruturas globais de disseminação de dados sobre biodiversidade na Web, esses dados ainda encontram-se em formatos que não permitem uma rápida colaboração com outras áreas do conhecimento humano. Este fato dificulta a descoberta de novos conhecimentos, por meio da análise dos relacionamentos entre as entidades presentes nos dados, além de não permitir uma rápida compreensão transdisciplinar dos fatos que as descrevem com fatos de outras áreas do conhecimento humano.

Além disso, a área de IB ainda encontra-se escassa de trabalhos relacionados ao desenvolvimento de ontologias e emprego das tecnologias da Web Semântica (WALLS *et al.*, 2014). Este fato gera um obstáculo para a publicação efetiva de dados sobre biodiversidade na WD.

No entanto, embora pareça apenas uma simples conversão de formato de dados (CSV para o modelo RDF), este processo não é trivial, pois deve envolver o uso compartilhado de conceitos ontológicos, além do compartilhamento de URI sobre as entidades biológicas conhecidas por todos os conjuntos de dados, ou seja, cada entidade biológica deverá ser vista como um único recurso Web (URI) por todos os dados que as referenciam, criando, dessa maneira, interoperabilidade entre os sistemas que utilizaram estes dados.

No Capítulo 4 foi apresentada a BioDSL, uma nova abordagem para publicação, vinculação e processamento por máquinas de dados sobre biodiversidade. Apesar de ser uma solução voltada para a área de biodiversidade, seus usuários (com pouco expertise na área de Informática e Web Semântica) ainda necessitam realizar configurações manualmente, para então executar os mapeamentos de dados para o modelo RDF. Além do mais, os usuários deverão instalar e configurar *endpoints* SPARQL para terem seus dados, no modelo RDF, disponíveis para consultas SPARQL na WD.

Neste contexto, julga-se necessário oferecer aos usuários de dados sobre biodiversidade uma infraestrutura semântica, baseada em tecnologias Web, para dar suporte à publicação, vinculação, consulta e integração de dados sobre biodiversidade na WD. Por meio desta infraestrutura semântica, os usuários poderão desenvolver seus *scripts* BioDSL e explorar as tecnologias da Web Semântica em um ambiente Web transparente, sem a necessidade de instalações e configurações adicionais de outros softwares.

Este capítulo está dividido em 8 seções. A seção 5.2 aborda os trabalhos do grupo de pesquisa que se relacionam ao tema desta tese, como: a extensão da OntoBio (seção 5.3); a busca semântica (seção 5.4) e o SWI Gazer (seção 5.5). A seção 5.6 apresenta a infraestrutura semântica proposta nesta tese, seu funcionamento e os principais componentes. A seção 5.7 apresenta experimentos realizados para demonstrar a integração de dados utilizando a Infraestrutura Semântica proposta nesta tese. A seção 5.8 apresenta conclusões sobre a Infraestrutura Semântica.

5.2 Trabalhos do Grupo de Pesquisa

As tecnologias da Web Semântica favorecem a interoperabilidade de dados científicos e não científicos, que podem ser a base para futuras pesquisas transdisciplinares. Pesquisas transdisciplinares sobre biodiversidade necessitam responder questões complexas, que podem envolver análises dos fatos conhecidos sobre entidades biológica, presentes nos dados sobre biodiversidade, com fatos, também conhecidos, sobre estas mesmas entidades, em dados de outras áreas do conhecimento humano, como os dados disponíveis na WD.

Nesse sentido, foi iniciada uma parceria entre o grupo do professor Dr. Dilvan de Abreu Moreira do Laboratório Intermídia do Instituto de Ciências Matemática e de Computação (ICMC) da Universidade de São Paulo (USP) com o Laboratório de Interoperabilidade Semântica (LIS) do Instituto Nacional de Pesquisas da Amazônia (INPA).

O LIS/INPA desenvolve a Ontologia de Biodiversidade – OntoBio (apresentada na seção 3.5) e trabalha em parceria com os programas e projetos científicos sobre biodiversidade do INPA, apresentados na seção 2.6.

A parceria entre o grupo de pesquisa do ICMC/USP com o LIS/INPA resultaram os trabalhos: a extensão da OntoBio (apresentada na seção 5.3); o desenvolvimento de uma arquitetura para sistemas de busca semântica em repositórios de biodiversidade (apresentada na seção 5.4); o desenvolvimento de um *gazetteer* semântico, o SWI *Gazetteer* (apresentado na seção 5.5) e o desenvolvimento da proposta desta tese, uma infraestrutura semântica para dados sobre biodiversidade (apresentada na seção 5.6).

5.3 A extensão da OntoBio

Este trabalho concentra-se na construção de uma infraestrutura semântica para dados sobre biodiversidade, logo, será dependente do uso de ontologias. De modo geral, uma ontologia estrutura o processo de aquisição do conhecimento, permitindo a compreensão e entendimento por máquinas no ambiente da Web Semântica. O sucesso de uma integração semântica de dados é dependente do poder de expressividade das ontologias utilizadas no mapeamento dos dados.

Não obstante, existe uma ausência de ontologias de domínio sofisticadas para biodiversidade, que inviabiliza o processo de descoberta e raciocínio sobre os dados. Tais ontologias estão começando a surgir, como a OntoBio, porém continuam a ser uma rara exceção e não a norma em pesquisas da IB (LAPP *et al.*, 2011; ALBUQUERQUE, 2011; SANTOS *et al.*, 2011).

A OntoBio é uma ontologia de domínio para biodiversidade desenvolvida pelo LIS/INPA, que foi apresentada na seção 3.5 (SANTOS *et al.*, 2011; GUIZZARDI, 2005). A OntoBio é baseada nos princípios da engenharia de ontologias e foi fundamentalmente baseada na pesquisa de (GUIZZARDI, 2005).

No entanto, para a realização dos experimentos relacionados com a busca semântica, foi necessário estender e criar uma nova versão desta ontologia, com base nos princípios do LD (AMANQUI *et al.*, 2013; AMANQUI *et al.*, 2013). A OntoBio estendida reutiliza URIs de outras ontologias conhecidas na WD, como *Darwin Core*, *Basic Formal Ontology*¹ (BFO), *Environment Ontology*² (ENVO), *Phenotype And Trait Ontology*³ (PATO), *Basic RDF Geo Vocabulary*⁴ (WGS84), *Geonames Ontology*⁵ e *DBpedia Ontology*⁶.

A reutilização de termos de outras ontologias (URIs) nos mapeamentos de dados sobre biodiversidade facilita a combinação destes dados com outros dados que utilizam os mesmo termos na WD. Dessa forma, cria-se uma integração automática dos dados sobre biodiversidade do INPA com outros conjuntos de dados na WD.

A versão estendida a BioDSL está disponível no site do BioPortal⁷ e a versão original é mantida no site da OntoBio⁸ do INPA.

5.4 Busca Semântica

Atualmente, formas tradicionais de pesquisa, baseadas em busca por palavras-chave, possuem baixa precisão e revocação, quando utilizadas em grandes volumes de dados sobre biodiversidade (AMANQUI, 2014). A precisão está relacionada a quantidade de elementos retornados pela busca com relevância a intensão do usuário. Já a revocação, está relacionada ao total de elementos relevantes a intensão do usuário e que foram recuperados pela busca.

Neste contexto, o grupo de pesquisa do ICMC/USP desenvolveu, em parceria com o LIS/INPA, uma arquitetura de busca semântica para recuperação de informações em repositórios de biodiversidade. Nesta arquitetura foram utilizados formatos e ferramentas da Web Semântica para representar os dados.

O projeto utilizou conjuntos de dados reais sobre biodiversidade cedidos pelo LIS/INPA, que incluíam dados sobre espécimes coletados (espécie, quando, onde foi a coleta, quem coletou, qual metodologia utilizada, etc.), informações geográficas (como latitude e longitude do local da coleta) e os ecossistemas onde foram realizadas as coletas.

Foi criado um mapeamento do conjunto de dados para o modelo RDF, utilizando-se a extensão da OntoBio, apresentada na seção 5.3.

Para a validação da arquitetura foram realizados experimentos com base em estudos de

¹ <https://github.com/BFO-ontology/BFO>

² <https://github.com/EnvironmentOntology/envo>

³ <https://github.com/pato-ontology/pato>

⁴ <https://www.w3.org/2003/01/geo/>

⁵ <http://www.geonames.org/ontology/documentation.html>

⁶ <http://wiki.dbpedia.org/services-resources/ontology>

⁷ <http://biportal.bioontology.org/projects/OntoBio>

⁸ <http://ontobio.inpa.gov.br/>

caso elaborados em conjunto com especialistas sobre biodiversidade do INPA. Os experimentos evidenciaram que a abordagem da busca semântica obteve um aumento de 27.8% no que diz respeito à precisão e de 25.1% na revocação em comparação com a busca por palavra-chave.

5.5 O SWI Gazetteer

Atualmente, repositórios de dados sobre biodiversidade na web vem contribuindo para a disseminação de dados sobre as coletas e observações de biodiversidade realizadas ao redor de todo o planeta. No entanto, grande parte destes dados possuem informações geográficas imprecisas ou ausentes (YESSON *et al.*, 2007; CARDOSO *et al.*, 2014; CARDOSO *et al.*, 2015; CARDOSO, 2015).

Os diretórios geográficos, também conhecidos como *Gazetteers*, associam nomes de lugares à coordenadas geográficas. No entanto, implementações atuais de *Gazetteers* não utilizam informações semânticas para responder consultas complexas, boa parte dos *Gazetteers* são apenas um diretório de nomes de lugares. Poucos *Gazetteers* disponibilizam suas informações associadas a ontologias de domínio para a desambiguação de nomes de lugares.

Com a necessidade de melhorar as informações geográficas contidas nos dados sobre biodiversidade, disponíveis em repositórios online (como o GBIF e *speciesLink*), o grupo de pesquisa do ICMC/USP desenvolveu um *Gazetteer* com suporte a tecnologias da Web Semântica (CARDOSO *et al.*, 2014; CARDOSO *et al.*, 2015). O *Semantic Web Interactive Gazetteer*, ou *SWI Gazetteer*, usa tecnologias da Web Semântica e técnicas de Recuperação de Informação Geográfica para associar coordenadas geográficas a nomes de lugares. O *SWI Gazetteer* também permite a disseminação de suas informações, usando formatos legíveis por máquinas, como o RDF.

Os resultados dos experimentos mostram que o *SWI Gazetteer* é capaz de aumentar o número de registros com coordenadas geográficas em amostras representativas de repositórios de dados sobre biodiversidade (como o GBIF e *speciesLink*). Em um dos experimentos, o *SWI Gazetteer* conseguiu adicionar coordenadas geográficas a cerca de 30.000 registros, aumentando a quantidade de registros com coordenadas de 30,29% para 57,5% do total de registros na amostra (que representou um aumento de 90% em relação aos registros com coordenadas inicial).

5.6 O funcionamento da Infraestrutura proposta

Esta tese propõem o desenvolvimento de uma infraestrutura semântica para dados sobre Biodiversidade com a finalidade de permitir que os seus usuários, especialistas em biodiversidade, criem bases de conhecimentos vinculadas com a WD a partir de dados tabulados sobre biodiversidade.

A etapa inicial do desenvolvimento deste projeto foi o estudo do problema, analisando o fluxo de dados e como as informações sobre biodiversidade são disseminados na Web, pelos programas e projetos científicos do INPA e parceiros do LIS, apresentados na seção 2.6.

A ideia principal da infraestrutura proposta é aplicar uma camada semântica sobre os dados sobre biodiversidade, presentes na Web, funcionando em paralelo, sem substituir as ferramentas, protocolos e padrões de dados e metadados desenvolvidos e mantidos pela IB.

Analisando os dados abertos sobre biodiversidade disseminados por redes, como *species-Link* e GBIF, foi detectado que boa parte dos dados são disseminados em formatos de arquivos de texto tabulado, como o CSV. Estes arquivos podem estar associados a arquivos de metadados, como no padrão DwC-A (seção 2.2.1).

A segunda fase foi verificar a viabilidade da adoção de uma ferramenta de mapeamento de dados tabulados para o modelo RDF, seguindo os princípios de LD. No entanto, foram detectadas algumas desvantagens nas ferramentas, apresentadas na seção 3.4.

Em resposta a estas desvantagens, foi desenvolvida uma nova abordagem para o mapeamento de dados tabulados sobre biodiversidade para o modelo RDF, denominada BioDSL, apresentada no Capítulo 4.

Para garantir a interoperabilidade, entre os conjuntos de dados sobre biodiversidade com outros conjuntos de dados na WD, foi desenvolvida uma estrutura de busca semântica, baseada nos trabalhos sobre busca semântica do grupo ICMC/USP, apresentados na seção 5.4.

A Busca Semântica, implementada na BioDSL, é focada na recuperação de URIs de entidades biológicas (espécies, lugares, instituições, etc.) para serem reutilizadas nos mapeamentos dos usuários. A Busca Semântica realiza a vinculação e a interoperabilidade entre os conjuntos de dados durante o processo de mapeamento para o modelo RDF.

Já a Busca Semântica, descrita na seção 5.4, foi implementada com o foco na comparação entre as técnicas de busca por palavra-chave e busca semântica. Essa pesquisa utilizou um conjunto de dados que foi mapeado para o modelo RDF, utilizando uma versão estendida da OntoBio (seção 5.4). No entanto, a arquitetura não permite que os usuários realizem alterações nos mapeamentos e na forma como a busca semântica será conduzida, além disso a interoperabilidade acontece apenas no nível dos termos compartilhados (URIs) da OntoBio estendida com os termos de ontologias da WD. A interoperabilidade não contempla as entidades biológicas (URI das entidades), descritas nos dados sobre biodiversidade.

Durante o desenvolvimento da BioDSL, foram detectadas algumas possíveis desvantagens em relação a pouca experiência dos usuários, biólogos em geral, na configuração do ambiente computacional necessário para rodar os *scripts* de mapeamentos da BioDSL. Por exemplo, o arquivo JAR da BioDSL, que interpreta os *scripts* BioDSL do usuário, necessita de um ambiente computacional com uma JVM configurada, um editor de texto para editar os *scripts* BioDSL e da execução de comandos específicos no terminal do Sistema Operacional,

Fonte: Elaborada pelo autor.

Figura 25 – Estrutura de funcionamento da infraestrutura semântica para dados sobre Biodiversidade

como descrito na seção 4.8. Além destas configurações, o usuário deverá ter que instalar e configurar *endpoints* SPARQL, caso queira realizar consultas SPARQL nos RDF resultantes dos seus mapeamentos.

Para atenuar estas possíveis desvantagens, foi desenvolvido um protótipo Web funcional, implementado em Java com o *Google Web Toolkit*⁹ (GWT), com o intuito de tornar transparente ao usuário, quando possível, todo o processo de mapeamento de dados, por meio de uma interface de *Rich Internet Application* (RIA) ou Aplicações de Internet Rica, denominada de BioDSL Web Editor, um *wrapper* para a linguagem BioDSL.

A Figura 25 apresenta a estrutura de funcionamento da infraestrutura semântica para dados sobre biodiversidade proposta nesta tese.

Na Figura 25, a área em verde claro representa a Web, o meio de comunicação entre elementos internos e externos da infraestrutura. O elemento cliente é representado pelo retângulo vermelho, onde os usuários da infraestrutura interagem com os dados e outros elementos via Web. O retângulo verde escuro, representa os repositórios de dados sobre biodiversidade, como o *speciesLink* e GBIF. A WD é representada pelo retângulo azul, no qual a infraestrutura explora os *endpoints* SPARQL nas buscas semânticas da BioDSL. Os usuários podem combinar consultas SPARQL entre os *endpoints* SPARQL da infraestrutura e da WD.

A sequência do fluxo das informações é demonstrada pelas setas numeradas na cor laranja da Figura 5.6:

- **Seta 1** – os usuários consultam dados sobre biodiversidade nos repositórios Web, como

⁹ <http://www.gwtproject.org/>

speciesLink e GBIF.

- **Seta 2** – os usuários carregam seus dados em arquivos CSV para o repositório de arquivos Web da infraestrutura.
- **Seta 3** – os usuários acessam a interface Web do BioDSL Web Editor, um *wrapper* da Linguagem BioDSL, para criar mapeamentos por meio de *scripts* BioDSL.
- **Seta 4** – ao executar os *scripts* BioDSL, o motor da BioDSL acessa o Repositório de Arquivos para manipular os arquivos CSV, as ontologias e os arquivos RDF resultantes dos mapeamentos.
- **Seta 5** – a BioDSL acessa os *endpoints* SPARQL da infraestrutura, para carregar os arquivos RDF do mapeamento e para realizar buscas semânticas nos conjuntos de dados mantidos pela infraestrutura.
- **Seta 6** – a BioDSL acessa *endpoints* SPARQL, externos a infraestrutura, pertencentes a WD, para realizar buscas semânticas.
- **Seta 7** – os usuários e agentes de softwares acessam, via Web, os *endpoints* SPARQL da infraestrutura.
- **Seta 8** – os usuários combinam suas consultas SPARQL entre os *endpoints* SPARQL da WD com os da Infraestrutura (buscando novos fatos sobre as entidades biológicas).

A Infraestrutura semântica é representada pelo retângulo com fundo de cor cinza. Os retângulos em azul claro, dentro do retângulo da cor cinza, representam os componentes da infraestrutura semântica. A BioDSL é o componente principal desta Infraestrutura. O interpretador da BioDSL fica na parte do servidor da infraestrutura semântica e interage com os demais componentes, como o repositório de arquivos, os *endpoints* SPARQL e GeoSPARQL da infraestrutura e da WD (externos a infraestrutura). Tais componentes são descritos a seguir:

Repositório de Arquivos – é um repositório Web para armazenamento dos arquivos CSV, *scripts* BioDSL, ontologias e arquivos dos modelos RDF e OWL resultantes dos mapeamentos. A BioDSL conecta-se ao repositório de arquivos da infraestrutura para armazenar os *scripts* BioDSL, guardar as ontologias utilizadas nos mapeamentos, ler os arquivos CSV e escrever em arquivos os modelos RDF e OWL resultantes dos mapeamentos. Os usuários poderão fazer *upload* dos seus arquivos CSV no repositório da infraestrutura, boa parte das funcionalidades deste repositório foram baseadas no trabalho desenvolvido em (SERIQUE; SANTOS; ALBUQUERQUE, 2014). A Figura 26 apresenta duas telas do repositório de arquivos. Na tela à esquerda, o usuário poderá ver a listagem de arquivos CSV e carregar novos arquivos pelo botão do canto direito inferior. A tela à direita, apresenta a caixa de seleção de arquivos CSV, com um arquivo selecionado pelo usuário. Este *upload* suporta apenas arquivos de texto com a extensão CSV.

Fonte: Elaborada pelo autor.

Figura 26 – Repositório de Arquivos da Infraestrutura

endpoints SPARQL e GeoSPARQL – a infraestrutura conta com *endpoints* SPARQL e GeoSPARQL para armazenar, realizar consultas e buscas semânticas entre os conjuntos de dados mapeados pelos usuários. Estes *endpoints* SPARQL e GeoSPARQL podem ser acessados por usuários externos a infraestrutura, ou seja, os dados dos usuários ficam, também, disponíveis na WD. A implementação atual da infraestrutura possui o *triplestore* Blazegraph, com uma extensão desenvolvida para melhorar as buscas semânticas da BioDSL.

BioDSL Web Editor – é o componente da infraestrutura que oferece aos usuários uma interface baseada em Aplicações de Internet Rica (*Rich Internet Application* – RIA) implementada em GWT, proporcionando um ambiente interativo para o desenvolvimento de *scripts* BioDSL. A Figura 27 apresenta alguns recursos do BioDSL Web Editor delimitados por uma linha tracejada de cor vermelha, associada a um círculo numerado.

- **Círculo 1** – mostra a área de edição de *scripts* BioDSL, delimitada por uma linha tracejada de cor vermelha;
- **Círculo 2** – apresenta uma área com abas para gerenciamento de ontologias, prefixos de ontologias, configurações do editor de *scripts* BioDSL e configurações para o visualizador de arquivos CSV.
- **Círculo 3** – mostra um console de informações sobre as ações que o usuário realiza na interface, serve também para apresentar *feedbacks* de erros ocorridos na execução do *script* BioDSL.
- **Círculo 4** – apresenta a área de edição e visualização dos arquivos CSV, caso um erro seja encontrado, como no tipos de dados, o usuário poderá realizar modificações nos arquivos CSV por meio desta interface Web.

Fonte: Elaborada pelo autor.

Figura 27 – Visão geral do BioDSL Web Editor

- **Círculo 5** - apresenta a área de visualização do RDF resultante dos mapeamentos, seguindo a sintaxe *turtle* do RDF.

O BioDSL Web Editor possui uma interface de desenvolvimento de *scripts* BioDSL, apresentada no círculo 1 da Figura 27. Esta ferramenta utiliza o ACE¹⁰, um editor Web de códigos escrito em *JavaScript*, integrado ao GWT para explorar os recursos de complementação de código e de mensagens informativas nas linhas de códigos com erros, como apresentado na Figura 28.

A complementação de código no BioDSL Web Editor é dinâmica, por exemplo, a cada ontologia adicionada, todos os seus termos são adicionados ao mecanismo de complementação de código do ACE. A complementação de código, também, ocorre com o nome das colunas do arquivo CSV e para as ontologias e prefixos adicionados. Quando o usuário tiver criando seus mapeamentos, poderá pesquisar pelos termos das ontologias, nome das colunas do arquivo CSV e prefixos, utilizando a complementação de código, como demonstrado na Figura 28. Por meio deste recurso, o usuário poderá explorar os modelos para uso das funções da Linguagem BioDSL, como *uri* e *map* (apresentadas nas respectivas seções 4.5.5 e 4.5.7).

Outro recurso explorado no ACE é o destaque de mensagens informativas, nas linhas de código com erros, apresentado na Figura 28. Com este recursos, pode-se apresentar *feedbacks* interativos aos usuários, melhorando a resolução de erros e outros problemas nos *scripts* BioDSL desenvolvidos pelos usuários.

¹⁰ <https://ace.c9.io/>

Fonte: Elaborada pelo autor.

Figura 32 – Aba de opções do editor de código BioDSL e de opções da Tabela

Codigo IBGE	Nome do Municipio	Codigo UF	UF	Estado	Latitude
☐ 5200050	Abadia de Goiás	52	GO	Goiás	-16.197
☐ 3100104	Abadia dos Dourados	31	MG	Minas Gerais	-18.4831
☐ 5200100	Abadiânia	52	GO	Goiás	-16.7573
☒ 3100203	Abatã	31	MG	Minas Gerais	-19.1551
☐ 1500107	Abatetuba	15	PA	Pará	-1.72183
☐ 2300101	Abaiara	23	CE	Ceará	-7.34588
☐ 2900108	Abaira	29	BA	Bahia	-13.2488
☐ 2900207	Abaré	29	BA	Bahia	-8.72073

Fonte: Elaborada pelo autor.

Figura 33 – Visualizador de arquivos CSV do BioDSL Web Editor

defineType da Linguagem BioDSL, seção 4.5.1. Caso um erro de tipo ocorra, o BioDSL Web Editor irá selecionar as linhas em que ocorreu o erro no arquivo CSV, o usuário poderá modificar os valores e permitir que o mapeamento ocorra sem problemas, como apresentado na tela de baixo na Figura 33.

Para permitir que os usuários visualizem o resultado em RDF dos mapeamento, o BioDSL Web Editor apresenta um RDF na sintaxe *turtle* para o usuário conferir o resultado do seu mapeamento. O círculo 5 apresenta área onde o RDF é apresentado ao usuário. Esta funcionalidade conta com uma opção de baixar o arquivo RDF e o arquivo OWL do mapeamento, sendo que o OWL é o resultado da junção das triplas do mapeamento com as triplas das ontologias carregados pelo usuário. O usuário poderá realizar consultas SPARQL nos dados resultantes por meio da opção carregar para *endpoint* SPARQL. A Figura 34 apresenta a interface de consultas SPARQL da infraestrutura.

Fonte: Elaborada pelo autor.

Figura 34 – Interface de consulta SPARQL da Infraestrutura

5.7 Experimentos com a Infraestrutura

As funcionalidades da infraestrutura semântica foram testadas, utilizando dados abertos sobre biodiversidade de instituições brasileiras, disponíveis nos repositórios do GBIF, *speciesLink* e SiBBR.

Apesar da facilidade de uso da BioDSL e do ambiente semântico, ainda não é possível que um biólogo, sem conhecimentos básicos de computação, use o sistema diretamente. Os usuários desses sistemas são os técnicos em informática (ou programadores) que trabalham assessorando grupos de pesquisa em biodiversidade. Por isso, a ajuda que tivemos de um grupo de usuários, especialistas em Web semântica e biodiversidade, do LIS/INPA foi muito importante. Eles trabalham com ontologias, dão suporte computacional às pesquisas sobre biodiversidade do INPA e contribuíram no desenvolvimento da linguagem BioDSL e da infraestrutura semântica.

Para demonstrar a capacidade de integração de dados sobre biodiversidade, realizou-se um experimento interligando os dados sobre espécies ameaçadas brasileiras, os dados sobre espécies da *DBpedia* e os dados do GBIF *Backbone Taxonomy* (GBIF Secretariat, 2017), apresentados na seção 5.7.1.

5.7.1 Integrando a lista de espécies ameaçadas brasileiras

O Instituto Chico Mendes de Conservação da Biodiversidade (ICMBIO) concluiu em 2014 a avaliação nacional do risco de extinção da fauna brasileira. Durante quatro anos, foram avaliados 12.256 táxons (8.924 de animais vertebrados e 3.332 de animais invertebrados). Os

resultados apontaram que, no Brasil, existem 1.173 táxons ameaçados de extinção, sendo 110 mamíferos, 234 aves, 80 répteis, 41 anfíbios, 353 peixes ósseos (310 água doce e 43 marinhos), 55 peixes cartilaginosos (54 marinhos e 1 água doce), 1 peixe-bruxa e 299 invertebrados. Totalizando-se 448 espécies Vulneráveis (VU), 406 em Perigo (EN), 318 Criticamente em Perigo (CR) e 1 Extinta na Natureza (EW). A lista com os 1.173 táxons pode ser baixada em formato de planilha eletrônica, disponível no site do ICMBIO.

Por outro lado, a *DBpedia* é um projeto com o objetivo de extrair conteúdo estruturado das informações das páginas da Wikipédia e disponibilizá-los na Web por meio de arquivos (no formato RDF) e de *endpoints* SPARQL. A *DBpedia* mantém URIs para entidades descritas nas páginas da Wikipédia. Estas URIs são muito utilizadas e conhecidas na WD.

A espécie é a entidade principal da lista de espécies ameaçadas do ICMBIO, sendo que o valor chave para essa entidade é a nomenclatura binomial, ou nomenclatura binária. A nomenclatura binomial é caracterizada pelo uso de dois nomes para definir uma espécie. Esses nomes são a junção do nome do gênero e de um nome restritivo específico (que serve como um adjetivo qualificativo para o gênero) que identificam uma espécie distinta. Esses nomes são importantes, pois permitem que pessoas em todo o mundo se comuniquem de forma inequívoca sobre as espécies.

O GBIF disponibiliza um conjunto de dados denominado GBIF *Backbone Taxonomy* (GBIF Secretariat, 2017), com o objetivo de reunir todos os nomes de espécies disseminados pelo seu portal. Esse conjunto de dados está no formato DwC-A e contém aproximadamente 5.718.729 registros, com 3.038.169 nomes e 2.211.443 sinônimos para espécies.

O objetivo final do experimento é produzir um conjunto de dados, em RDF, sobre as espécies do ICMBIO em risco de extinção interligando as informações das 3 fontes de dados (ICMBIO, DBpedia e GBIF).

A primeira parte do experimento consiste na interligação da lista de espécies ameaçadas com os dados da DBpedia. As triplas (S,P,O) sobre espécies da *DBpedia* possuem uma propriedade (P) para o valor binomial, definida pela URI `<http://dbpedia.org/property/binomial>`. A planilha sobre espécies ameaçadas de extinção do ICMBIO possui uma coluna denominada "Espécie", contendo a nomenclatura binomial do registro da espécie ameaçada. Alguns registros possuem, também, informações sobre a autoria da espécie, no entanto, para realizar este experimento foi necessário isolar apenas a nomenclatura binomial em uma nova coluna, denominada "binomial".

5.7.1.1 Conversão dos dados do ICMBIO e integração com a DBpedia

Na primeira parte deste experimento, a infraestrutura semântica converteu os dados sobre espécies ameaçadas da fauna brasileira do ICMBIO para o formato RDF e os integrou com dados da *DBpedia*. Foi utilizado um *script* escrito na linguagem BioDSL, para conversão, juntamente

com a função *matching*, para recuperar e reutilizar as URIs de entidades de espécies das triplas da *DBpedia*. As espécies do ICMBIO usam agora as URIs da *DBpedia* para os sujeitos (S) das triplas resultantes do mapeamento para o modelo RDF. Usando-se essas URIs é possível recuperar informações sobre as espécies na *DBpedia* e em dezenas de repositórios que usam as URIs da *DBpedia*. Ao final do processo, as triplas resultantes do mapeamento foram carregadas em uma instância do *triplestore Blazegraph*.

Para este experimento, foi realizado apenas um mapeamento formando duas triplas para cada espécie. As triplas contém o (S) proveniente da URI recuperada na *DBpedia*. A primeira tripla, linha 11 do código fonte 8, possui uma propriedade *type* do RDF vinculada à URI `<http://dbpedia.org/ontology/Species>`, que é um termo da ontologia da *DBpedia* correspondente às espécies. A segunda tripla contém uma propriedade (P) com o *label* (rótulo) do RDF-S e o objeto (O) com o valor binomial proveniente da planilha do ICMBIO, como mostra a linha 12 do código fonte 8.

Código-fonte 8 – Mapeamento de espécies brasileiras ameaçadas

```
1: uriBase "http://java.icmc.usp.br/"
2: uri (spe, "species", csv.binomial)
3: matching (speDbpedia) {
4:     endpoint "http://java.icmc.usp.br:1680/speciesdbpedia/
      namespace/kb/sparql"
5:     vendor BLAZEGRAPH
6:     search csv.binomial
7:     predicate "http://dbpedia.org/property/binomial"
8:     defaultURI spe
9: }
10: map (species, speDbpedia) {
11:     rdf.type "http://dbpedia.org/ontology/Species"
12:     rdfs.label csv.binomial
13: }
```

Na linha 3 do código fonte 8, uma função *matching* foi declarada para recuperar as URIs das espécies conhecidas na *DBpedia*. A função irá explorar o *endpoint* da *DBpedia* (linha 4) para recuperar as URIs (S) que possuam propriedades (P) `<http://dbpedia.org/property/binomial>` (linha 7) com valores de objeto (O) correspondentes ao valor binomial da planilha. Caso uma URI não seja encontrada na *DBpedia*, o objeto *spe*, declarado pela função *uri* na linha 2, será usado.

Os dados da *DBpedia* totalizam 225.909 URIs (S), contendo o predicado binomial `<http://dbpedia.org/property/binomial>`, todavia, apenas 497 URIs são correspondentes aos registros de espécies brasileiras em extinção. Com o BioDSL *script*, apresentado no código fonte

8, foi possível recuperar as 497 URIs da *DBpedia*, sendo que 493 URIs foram recuperadas com 100% de grau de comparação entre os valores binomiais. 4 registros possuíam problemas na comparação, 3 registros tiveram um carácter trocado (2 com espaços em branco a mais e 1 com um carácter errado) e apenas 1 registro teve uma correspondência parcial (com apenas o valor de gênero correto).

5.7.1.2 Conversão dos dados do GBIF Backbone Taxonomy

A segunda parte do experimento envolveu o uso do conjunto de dados GBIF *Backbone Taxonomy*, que objetiva cobrir os nomes de espécies utilizados pelo portal do GBIF (GBIF Secretariat, 2017). Esse conjunto de dados encontra-se no formato DwC-A, logo foi necessário seu mapeamento para o modelo RDF. O código fonte 9 apresenta o *script* BioDSL do mapeamento do GBIF *Backbone Taxonomy* para o modelo RDF.

Código-fonte 9 – Mapeamento de espécies brasileiras ameaçadas

```
1: uriBase "http://java.icmc.usp.br/"
2: uri (speciesURI, "species", csv.scientificName)
3: map (species, speciesURI) {
4:     rdfs.label csv.scientificName
5: }
```

Nos dados do GBIF, os nomes científicos das espécies estão na coluna “*scientificName*” (correspondente ao termo `<http://rs.tdwg.org/dwc/terms/scientificName>` do padrão DwC), que também foi utilizada na construção das URIs pela função *uri* da BioDSL, código fonte 9 linha 2. No total foram geradas 5.867.989 triplas resultantes do mapeamento realizado pela função *map*, linha 3. Ao final do mapeamento, as triplas foram carregadas para um *endpoint* SPARQL¹¹.

5.7.1.3 Interligação dos conjuntos de dados

O *script* BioDSL do código fonte 8 foi reexecutado, com o novo valor para o parâmetro *endpoint* (linha 3) com as triplas do GBIF *Backbone Taxonomy*. Dentre os 1173 registros de espécies ameaçadas, a BioDSL conseguiu recuperar 967 URIs com *matching* completo para o valor binomial das espécies ameaçadas (82,4%) e 149 URIs com *matching* parcial com apenas um dos nomes (12,7%). 36 URIs não tiveram nenhuma correspondência de nomes (3,1%), pois as consultas não retornaram registros, então a função *matching* utilizou a URI fornecida pelo objeto URI “*spe*”, linha 8, declarado na função *uri* da linha 2 do código fonte 8. O GBIF *Backbone Taxonomy* não possui registros para 21 (1,8%) nome presentes na lista espécies ameaçadas do ICMBIO.

¹¹ <http://java.icmc.usp.br:2280/species/namespace/GBIFBackboneTaxonomy/sparql>

O conjunto de dados final, em RDF, sobre as espécies do ICMBIO em risco de extinção tem informações sobre as 3 fontes de dados (ICMBIO, DBpedia e GBIF). Ele demonstra que, com pouco conhecimento sobre linguagens de programação e DSLs, um usuário técnico em informática pode integrar dados do seu laboratório com dados da LOD e de outros usuários da infraestrutura semântica, com poucas linhas de código da linguagem BioDSL.

5.7.2 Integrando dados das eleições de 2014

Uma das principais dificuldades no desenvolvimento deste trabalho, foi encontrar usuários especialistas em biodiversidade com domínio em computação (poucos programadores ou biólogos possuem esse perfil). Apesar da ajuda de colaboradores do LIS-INPA, esse fato dificultou os testes das funcionalidades da infraestrutura semântica, pois haviam poucos profissionais a nossa disposição para testes. Para contornar essa dificuldade, decidimos usar a flexibilidade da infraestrutura e da BioDSL para integrar dados de outros domínios (que tivessem formatos de dados parecidos com os usados para biodiversidade).

O nosso grupo, no ICMC/USP, tem alunos desenvolvendo um trabalho denominado Disponibilização de Informações sobre Campanhas Eleitorais em Formato Aberto (*Open Data*) na Web. Esse trabalho visa transformar dados sobre eleições e políticos brasileiros, disponíveis em diversos formatos tabulares (CSV), em triplas RDF disponíveis num *endpoint* SPARQL. Os dados principais vêm do Tribunal Superior Eleitoral (TSE), que disponibiliza, por meio de um Web site¹², dados sobre as eleições brasileiras desde o ano de 1945. Todos os dados do TSE estão disponíveis em formato CSV e podem ser importados para qualquer programa, base de dados ou planilha eletrônica.

A ideia é disponibilizar os dados do TSE, sobre as eleições e políticos, na LOD e adicionar informações de outros sites sobre os políticos (processos na justiça, leis votadas, etc). A BioDSL está sendo usada nesse projeto para converter os dados para RDF e conecta-los a LOD (usando URIs da DBpedia).

Para demonstrar a capacidade de conectar informações de um domínio de conhecimento para entidades na LOD, foi feito um experimento de integração de dados sobre as eleições de 2014. Alunos participantes do projeto executaram esse experimento de integração utilizando a infraestrutura semântica e a linguagem BioDSL. Este experimento consistiu na conversão dos dados CSV da eleição de 2014 para RDF reusando as URIs de fontes conhecidas na WD (*DBpedia* e a *Wikidata*) para criar ligações entre os dados da eleição e a WD.

O *endpoint* SPARQL da *DBpedia* em português¹³ possui 21.432 URIs sobre políticos (URIs que pertencem a classe `<http://dbpedia.org/ontology/Politician>`). No total, foram mapeados, dos dados do TSE, 3.663 candidatos pertencentes ao estado de São Paulo, que participaram das eleições de 2014. Deste total, foram encontrados 207 políticos com URIs na *DBpedia*.

¹² <http://www.tse.jus.br/hotSites/pesquisas-eleitorais/>

¹³ <http://pt.dbpedia.org/sparql>

Neste experimento, o valor chave nos dados do TSE foi o nome do político, uma vez que as demais informações, como o número do título de eleitor (que são únicas por natureza) não estão disponíveis na LOD.

Código-fonte 10 – Mapeamento de dados sobre as eleições paulistas de 2014

```
1: uriBase "http://purl.org/politics/"
2: uri (politicianURI, "politician", csv.nomeCandidato)
3: matching(politicianLOD){
4:     endpoint      "http://pt.dbpedia.org/sparql"
5:     vendor         VIRTUOSO
6:     search         csv.nomeCandidato
7:     type           "http://dbpedia.org/ontology/Politician"
8:     defaultURI     politicianURI
9: }
10: map (politician, politicianLOD) {
11:     rdfs.label     csv.nomeCandidato
12: }
```

O *script* BioDSL do código fonte 10 foi utilizado no experimento. Foram recuperados 207 URIs de políticos na *DBpedia* em português, sendo que 129 foram recuperadas com 100% de grau de comparação de *strings* e 78 com alguma penalidade na comparação, sendo 29 URIs recuperadas com diferença de apenas um caracter (letra), 7 URIs foram recuperadas com diferença de dois caracteres e 42 URIs foram recuperadas com mais de 2 caracteres de diferença (menor precisão).

Como era esperado, nem todos os candidatos das eleições de 2014 são conhecidos na *DBpedia* (só candidatos que possuem páginas na Wikipédia), consequentemente as URIs recuperadas eram relacionada aos nomes dos candidatos mais conhecidos publicamente (tiveram mandatos anteriores ou já eram personalidades bem conhecidas). Com base nas URIs, recuperadas e utilizadas no RDF do mapeamento, é possível associar novas informações da Wikipédia a estes políticos, aumentando a utilidade do conjunto de dados do TSE. Com pequenas modificações no *script* do código fonte 10 é possível fazer o mesmo para outros grupos de entidades relacionadas as eleições, como partido políticos, cidades, etc.

5.8 Considerações Finais

A integração de dados científicos permite uma compreensão transdisciplinar dos fatos que ocorrem em uma determinada pesquisa científica.

Tendo isso em mente, projetou-se uma infraestrutura genérica para integração de dados sobre biodiversidade, utilizando tecnologias da Web Semântica (como RDF, OWL, SPARQL e

GeoSPARQL) para realizar essa integração. A proposta demonstra que as tecnologias da Web Semântica são uma opção sofisticada para integração de dados e conhecimento compartilhado, por meio das ontologias.

A Linguagem BioDSL é uma solução genérica e pode ser utilizada em outros tipos de dados, ou seja, ela não se restringe apenas a área de biodiversidade.

Essa abordagem mostra uma evolução em relação a forma atual de disseminação de dados, baseada em padrões de dados e metadados com baixo poder de expressividade, que muitas vezes não são totalmente interoperáveis entre si.

A infraestrutura semântica possui a capacidade de ser expandida para outras áreas fora o escopo da biodiversidade. Essa infraestrutura semântica pode ser expandida e combinada com outras fontes de dados (de outras áreas do conhecimento) por meio da WD.

CONCLUSÃO E TRABALHOS FUTUROS

A hipótese desta tese foi: “*A interoperabilidade semântica irá propiciar que dados científicos sejam consultados e consumidos de forma transdisciplinar, no campo da biodiversidade ou em qualquer outro campo do conhecimento humano*”. Partindo desta hipótese, o objetivo geral foi desenvolver uma infraestrutura computacional para integração semântica de dados científicos sobre a biodiversidade. A infraestrutura semântica, desenvolvida nesta tese, baseia-se no uso das tecnologias da Web Semântica. Foi criado um ambiente colaborativo e uma linguagem de domínio específico, a BioDSL, juntamente a uma plataforma Web interativa para auxiliar o usuário a criar bases de conhecimentos em RDF integradas a outros conjuntos de dados e com a LOD. Como resultado, obteve-se um ambiente computacional, transdisciplinar, onde usuários podem interligar conjuntos de dados entre si e com a WD, para serem capazes de explorar questões científicas que extrapolem o domínio dos seus dados.

6.1 Resumo do trabalho realizado

No Capítulo 1 foram apresentados a contextualização dos aspectos transdisciplinares e pesquisas transformadoras. Em seguida, foi apresentada a área de Biodiversidade, que possui um caráter científico altamente transdisciplinar, e uma nova área, a IB, que é a aplicação de TI para solucionar problemas da Biodiversidade relacionados aos dados e informações (gestão, armazenamento, disseminação, etc.). Em um outro caminho, a área de Web Semântica, propõe a construção de uma WD, uma evolução da Web atual, integrando dados de diversas áreas do conhecimento humano em um 3G. Assim, um proeminente ambiente transdisciplinar começa a ser construído, por meio dos relacionamentos explícitos entre os recursos, entidades e conceitos sobre diversas disciplinas na Web. No entanto, nota-se uma pequena presença dos dados sobre Biodiversidade na WD, apesar de todos os avanços da IB na direção da disseminação global de dados sobre biodiversidade, como ocorrências de biodiversidade, registros de espécimes em museus, etc. A hipótese defendida nesta pesquisa se baseia na integração semântica de dados

científicos, como uma base para o desenvolvimento de futuras pesquisas transdisciplinares, incluindo, também, as pesquisas sobre Biodiversidade. O objetivo deste trabalho foi desenvolver uma infraestrutura semântica computacional, para dados sobre biodiversidade, integrando-os entre si e com a WD, permitindo que usuários realizem consultas e mapeamentos por meio de um ambiente Web interativo.

O Capítulo 2 apresentou o referencial teórico, norteador deste trabalho, sobre problemas de integração de dados abertos na área de IB, bem como os programas e projetos científicos, apoiados pelo nosso parceiro de pesquisa, o LIS/INPA. O referencial teórico apresentou os padrões de dados e metadados, os protocolos de disseminação de dados, as ferramentas para gestão de dados e os repositórios de disseminação de dados sobre biodiversidade via Web. Foram destacados os seguintes padrões de dados e metadados de biodiversidade: Darwin Core; ABCD e EML. Na sequência, foram apresentadas as seguintes ferramentas de gestão de informação sobre biodiversidade: BRAMS; Specify; Metacat e Morpho. Ainda, foram apresentados as evoluções dos protocolos de disseminação de dados sobre biodiversidade, tais como: DiGIR; BioCASE; TAPIR e IPT. Também foram apresentados, os repositórios de disseminação de dados sobre biodiversidade, tais como: SinBiota FAPESP; *speciesLink*; GBIF e SiBBR. Ainda foram apresentados os principais programas e projetos apoiados pelo nosso parceiro de pesquisa, LIS/INPA, com a intenção de identificar os padrões de dados e metadados, ferramentas e repositórios utilizados nas pesquisas sobre a biodiversidade. Foram destacados projetos e programas: o PPBio, o LBA, o PELD, o PDBFF e o TEAM. Uma análise sobre a utilização da IB nos projetos e programas supracitados foi realizada dando uma visão geral de como a IB é aplicada e de como os dados são produzidos e geridos por eles.

O Capítulo 3 apresentou, suplementarmente, o material de referencial teórico sobre as tecnologias da Web Semântica. Os conceitos relatados incluíram: a importância do uso de URI para identificar recursos na Web Semântica; os modelos de dados RDF e RDFS; a Linguagem de definição de Ontologias na Web – OWL; Os protocolos e linguagens de consulta em RDF (SPARQL e GeoSPARQL); além dos conceitos e recomendações do LD.

Tanto o Capítulo 2 quanto o Capítulo 3 apresentaram os materiais de referencial teórico sobre áreas distintas, a IB e a Web Semântica, que foram integradas para o desenvolvimento do objetivo deste trabalho, a construção de uma infraestrutura semântica para os dados sobre biodiversidade, e apresentadas no decorrer dos Capítulos 4 e 5 desta tese.

O Capítulo 4 apresentou a linguagem BioDSL. Os usuários de dados sobre biodiversidade podem, por meio desta simples linguagem declarativa, desenvolver pequenos *scripts* (programas) que definirão como serão as estruturas do RDF para descrever as entidades presentes nos dados sobre biodiversidade. Seguindo as recomendações da LD, o usuário poderá reutilizar URIs de entidades conhecidas em outros conjuntos de dados, por meio de uma busca semântica adaptada para realizar consultas em *endpoints* SPARQL, disponíveis tanto na infraestrutura semântica (apresentada no Capítulo 5) quanto na WD. A BioDSL conta, também, com funções

de validação de dados, onde os usuários poderão verificar se os dados do arquivo CSV estão de acordo com tipos primitivos de dados (*string*, numérico, data, hora, etc.). A nível semântico, a BioDSL permite que o usuário crie RDF, utilizando termos de ontologias em OWL. Além disso, o usuário poderá executar o OWL *reasoning* para verificar se o modelo RDF, resultante do seu mapeamento, está semanticamente correto com as definições das Ontologias que foram utilizadas no mapeamento. No final do processamento da BioDSL, o usuário obterá o seu modelo RDF em diversos formatos de arquivo RDF (RDF XML, turtle, N3, etc.), além de, se preferir, carregar o RDF diretamente em um *triplestore* SPARQL. Dessa forma, o usuário permite que outros usuários possam, também, realizar consultas nos dados e realizarem buscas semânticas para integrar suas entidades.

O Capítulo 5 apresentou as contribuições dos trabalhos do grupo relacionados com esta tese, além de apresentar o funcionamento da implementação do protótipo da infraestrutura semântica proposta. O grupo de pesquisas do ICMC/USP possui uma parceria com o LIS/INPA que resultou nos trabalhos de extensão da OntoBio, arquitetura de busca semântica em dados sobre biodiversidade, o SWI *Gazetteer* e a proposta deste trabalho. O funcionamento da infraestrutura semântica proposta tem a Linguagem BioDSL como o principal componente. Um ambiente com interface Web foi desenvolvido para a BioDSL com a intenção de automatizar ao máximo todo o processo de mapeamento de dados, evitando que o usuário tenha que realizar configurações em seus computadores para usufruírem da ferramenta. Esse ambiente foi denominado de BioDSL Web Editor e conta com um repositório de arquivos Web, uma interface Web para desenvolvimento de *scripts* BioDSL, um gerenciador de ontologias e prefixos de URI, um gerenciador de arquivos CSV, um visualizador de RDF e uma interface para consultas SPARQL. Esse capítulo também apresentou resultados de experimentos que demonstram a capacidade de integração de dados do ambiente.

6.2 Parcerias

O grupo de pesquisa do ICMC/USP e o LIS/INPA possuem uma parceria em diversos trabalhos anteriores, como a extensão da OntoBio, a arquitetura de busca semântica, o SWI *Gazetteer*. Neste trabalho foi realizado um estágio de nove meses (fevereiro de 2013 à outubro de 2013) onde o autor desta tese frequentou o INPA, na cidade de Manaus – Amazonas, desenvolvendo, realizando pesquisas sobre IB e levantamentos de requisitos com especialistas sobre biodiversidade do INPA (entrevistas, conversas informais, etc.).

Durante o período do estágio, foram desenvolvidas atividades como: o desenvolvimento de um software de repositório de arquivos na Web, denominado Mo Porã¹; publicação de um capítulo de livro sobre o software (SERIQUE; SANTOS; ALBUQUERQUE, 2014); e a redação do texto para a qualificação do doutorado.

¹ <http://cenarios.inpa.gov.br/>

6.3 Contribuições ao Estado da Arte

Como contribuições científicas alcançadas por esta pesquisa podemos elencar:

- O levantamento do estado da arte relacionado às pesquisas sobre a biodiversidade, realizadas em instituições brasileiras, é apresentado no Capítulo 2. A análise do fluxo de dados sobre biodiversidade, nas pesquisas amazônicas, demonstra como esses dados são gerados, curados e disseminados. Essas informações contribuíram no amadurecimento da proposta desta tese e na construção da infraestrutura semântica para dados sobre biodiversidade.
- A criação de um ambiente LD integrado necessita atender os princípios de publicação, apresentados na seção 3.3. No entanto, esta pesquisa constatou que as ferramentas de mapeamento de dados para o modelo RDF, apresentadas na seção 3.4, não contemplam os requisitos necessários para publicação efetiva na LOD. Elas oferecem pouco suporte para vincular dados com a WD de forma automática ou semi-automática, necessitando que seus usuários realizem outras atividades complementares para a publicação dos dados em LD.
- O desenvolvimento da Linguagem BioDSL, apresentado no Capítulo 4, contempla as desvantagens das linguagens de mapeamento (seção 3.4), oferecendo aos usuários um ambiente transparente, encobrendo a complexidade do processo de mapeamento baseado nos princípios da LD. Essa linguagem é uma DSL para mapeamento de dados sobre a biodiversidade (no formato CSV) para o modelo RDF. Seus recursos são descritos no artigo “*BioDSL: a domain-specific language for mapping and dissemination of biodiversity data in the LOD*” (SERIQUE *et al.*, 2016).
- A reutilização de URIs de entidades conhecidas na WD garante a interoperabilidade entre os conjuntos de dados sobre biodiversidade com outros conjuntos de dados na WD. Por isso, o desenvolvimento de mecanismos automatizados para a reutilização dessas URIs é muito importante. A função *matching*, da linguagem BioDSL, implementa esse mecanismo de busca semântica e é uma evolução de trabalhos desenvolvidos, pelo autor, no grupo de pesquisa do ICMC/USP descritos nos artigos “*Improving Biodiversity Data Retrieval through Semantic Search and Ontologies*” (AMANQUI *et al.*, 2014) e “*SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data*” (CARDOSO *et al.*, 2015). Na função *matching*, os usuários podem realizar configurações para definir semanticamente a intenção da sua busca, com uso de termos (URIs de ontologias) e dos dados provenientes dos arquivos CSV. Essa função explora dados contidos em *endpoints* SPARQL da infraestrutura semântica, da LOD e do SWI *Gazetteer* e recupera as URIs para utilizar no mapeamento, garantindo assim, interoperabilidade com ligações entre os diversos conjuntos de dados da WD.
- Os resultados dos experimentos, apresentados na seção 5.7, demonstram a capacidade de integração (por reuso de URIs) de dados locais à infraestrutura semântica, com dados

da WD (e de outros usuários). A infraestrutura reduz a complexidade envolvida nesse processo e a torna mais acessível aos usuários.

- Foi desenvolvido, também, o BioDSL Web Editor, um ambiente Web interativo para o desenvolvimento de *scripts* na linguagem BioDSL. Essa ferramenta serve como *front-end* para a infraestrutura semântica, proposta nesta tese. Ela tem a linguagem BioDSL como coração e conecta as funcionalidades e recursos (APIs, *triplestores*, arquivos RDF, ontologias em OWL, repositório de arquivos, etc.) da infraestrutura semântica. O repositório de arquivos, utilizado no BioDSL Web Editor, foi uma evolução do trabalho publicado no artigo “*Mo Porã: Uma ferramenta para o gerenciamento distribuído de repositórios de dados científicos na web*” (SERIQUE; SANTOS; ALBUQUERQUE, 2014). No experimento com dados das eleições paulistanas de 2014, seção 5.7.2, os usuários utilizaram o BioDSL Web Editor para desenvolver os *scripts* BioDSL, sem a necessidade de instalação e configuração do ambiente para o desenvolvimento dos *scripts* BioDSL (configuração de JVM, *triplestores*, etc.). Isso demonstra que a infraestrutura semântica, proposta nesta tese, além de contemplar o domínio da Biodiversidade também pode ser utilizada para integrar dados de outros domínios.

6.4 Publicações geradas até o momento da defesa

Foram geradas as seguintes publicações durante o desenvolvimento deste doutorado:

- O artigo “*BioDSL: a domain-specific language for mapping and dissemination of biodiversity data in the LOD*” foi apresentado no *X Brazilian e-Science Workshop* do XXXVI Congresso da Sociedade Brasileira de Computação (SERIQUE *et al.*, 2016).
- O artigo “*SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data. Future Generation Computer Systems*” foi publicado no periódico *Future Generation Computer Systems*, fator de impacto 3,997 (CARDOSO *et al.*, 2015).
- O artigo “*Improving Biodiversity Data Retrieval through Semantic Search and Ontologies*” foi apresentado no *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (AMANQUI *et al.*, 2014).
- O artigo “*A Gazetteer for Biodiversity Data as a Linked Open Data Solution*” foi apresentado no *2014 IEEE 23rd International WETICE Conference* (CARDOSO *et al.*, 2014).
- O capítulo de livro “*Mo Porã: Uma ferramenta para o gerenciamento distribuído de repositórios de dados científicos na web*” foi publicado no livro *Cenários* (SERIQUE; SANTOS; ALBUQUERQUE, 2014).

- O artigo “*Semantic Search Architecture for Retrieving Information in Biodiversity Repositories*” foi apresentado no *6th Seminar on Ontology Research in Brazil* (AMANQUI *et al.*, 2013).

6.5 Trabalhos futuros

A lista de trabalhos futuros:

- Como principal trabalho futuro, pretende-se lançar como software livre uma versão estável da linguagem BioDSL e dos protótipos da infraestrutura semântica propostas nesta tese;
- Pesquisar novos *frameworks* de desenvolvimento Web para a interface Web da Infraestrutura. O framework GWT, usado no projeto, tem problemas de lentidão na liberação de novas funcionalidades, além de ser muito laborioso a integração e manutenção de outras bibliotecas externas ao GWT;
- Tornar a BioDSL mais genérica, quanto possível, para poder ser utilizada por outras áreas do conhecimento, como ferramenta de publicação de dados na WD;
- Aperfeiçoar a Linguagem BioDSL com adição de novas funcionalidades e maior integração com o BioDSL Web Editor;
- Aperfeiçoar os mecanismos de busca semântica que contemplem todos os tipos de *endpoints* SPARQL e *Gazetteer* semânticos, incluindo *endpoints* SPARQL que não possuem o recurso de *full text search*;
- Criar um *hub* para desambiguação de nomes de entidades biológicas conhecidas (espécies, taxonomias, lugares, habitats, etc.) para facilitar a integração dos dados, centralizando todas as URIs conhecidas sobre as entidades em um único lugar (*endpoint* SPARQL) e melhorar os resultados para buscas semânticas e consultas federadas SPARQL;

REFERÊNCIAS

Access to Biological Collections Data task group. **Access to Biological Collection Data (ABCD), Version 2.06**. 2005. Citado nas páginas 30 e 44.

ALBUQUERQUE, A. C. F.; SANTOS, J. L. C. dos; de Castro Júnior, A. N. Elicitation Taxonomy for Acquiring Biodiversity Knowledge. In: _____. **Knowledge Engineering and Semantic Web: 7th International Conference, KESW 2016, Prague, Czech Republic, September 21-23, 2016, Proceedings**. Cham: Springer International Publishing, 2016. p. 157–172. ISBN 978-3-319-45880-9. Citado na página 95.

ALBUQUERQUE, A. C. F. A. **Um Framework Conceitual para Integrar Conhecimento Tácito Científico**. 178 p. Tese (thesis) — Universidade Federal do Amazonas, 2016. Citado nas páginas 32, 82, 95 e 96.

ALBUQUERQUE, A. C. F. A. C. F. **Desenvolvimento de uma Ontologia de Domínio para Modelagem de Biodiversidade**. 135 p. Tese (Doutorado) — UFAM, 2011. Citado nas páginas 30, 40, 44, 95, 96 e 123.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da Informação**, v. 32, n. 3, p. 7–20, dec 2003. ISSN 0100-1965. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S0100-19652003000300002{&}lng=pt{&}nrm>. Citado na página 82.

AMANQUI, F. K.; SERIQUE, K. J.; CARDOSO, S. D.; SANTOS, J. L. D.; ALBUQUERQUE, A.; MOREIRA, D. A. Improving Biodiversity Data Retrieval through Semantic Search and Ontologies. In: **2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)**. IEEE, 2014. p. 274–281. ISBN 978-1-4799-4143-8. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6927553>>. Citado nas páginas 36, 144 e 145.

AMANQUI, F. K. F. K.; SERIQUE, K. J. K. J.; LAMPING, F.; ALBUQUERQUE22, A. C. F.; Dos Santos, J. L. C.; MOREIRA, D. A. Semantic Search Architecture for Retrieving Information in Biodiversity Repositories. In: **ceur-ws.org**. [s.n.], 2013. v. 1041, p. 83–93. Disponível em: <http://ceur-ws.org/Vol-1041/ontobras-2013{_}paper29.> Citado nas páginas 36, 124 e 146.

AMANQUI, F. K. M. **Uma arquitetura para sistemas de busca semântica para recuperação de informações em repositórios de biodiversidade**. 82 p. Tese (master thesis) — Universidade de São Paulo, 2014. Citado na página 124.

AMANQUI, F. M.; SERIQUE, K. J.; LAMPING, F.; ALBUQUERQUE, A. C. F.; MOREIRA, D. A.; KARINA, F.; AMANQUI, M.; Campos dos Santos, J. L. Architecture for Semantic Search Systems for Retrieving Information in Biodiversity Repositories. **Anais do 28º Simpósio Brasileiro de Banco de Dados**, Brazilian Computer Society, Recife, p. 121–126, 2013. Disponível em: <<http://sbbd2013.cin.ufpe.br/Proceedings/application/01.html>>. Citado na página 124.

ANTEZANA, E.; KUIPER, M.; MIRONOV, V. Biological knowledge management: the emerging role of the Semantic Web technologies. **Briefings in bioinformatics**, v. 10, n. 4, p. 392–407, jul 2009. ISSN 1477-4054. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/19457869>>. Citado na página 34.

ARIÑO, A. H.; CHAVAN, V.; KING, N. The Biodiversity Informatics Potential Index. **BMC Bioinformatics**, v. 12, n. Suppl 15, p. S4, jan 2011. ISSN 1471-2105. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3287447&tool=pmcentrez&rendertype=ab>>. Citado na página 40.

BACH, K.; SCHÄFER, D.; ENKE, N.; SEEGER, B.; GEMEINHOLZER, B.; BENDIX, J. A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. **Ecological Informatics**, p. 16–24, 2012. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1574954111000987>>. Citado na página 41.

BATISTELLA, M.; ALVES, D.; ARTAXO, P.; BUSTAMANTE, M. M. C.; KELLER, M.; LUIZÃO, F. J.; MARENGO, J. A.; MARTINELLI, L.; NOBRE, C. A.; ALVES, M. D. S.; ANDREAE, M. O.; ASNER, G. P.; AVISSAR, R.; BECKER, K.; BUSTAMANTE, M. M. C.; COHEN, J.; MARIA, A. F.; DIAS, S.; SILVA, P. L.; FIGUEIREDO, R.; GALARRAGA-SANCHEZ, R. H.; KRUIJT, B.; KRUSCHE, A. V.; LLERENA, C. A.; LLOYD, J.; LUIZÃO, F. J.; MELACK, J.; MOUTINHO, R. D. S.; NOBRE, A. D.; NOVO, E. M. L. M.; PIEDADE, M. T. F. Plano Científico LBA2 – PROGRAMA DE PESQUISAS SOBRE INTERAÇÕES BIOSFERA-ATMOSFERA NA AMAZÔNIA. **INPA**, 2007. Citado nas páginas 60, 61, 63 e 64.

BATTLE, R.; KOLAS, D. Enabling the Geospatial SemanticWeb with Parliament and GeoSPARQL. **Semantic Web Journal**, v. 0, n. 0, p. 1–17, 2012. ISSN 1570-0844. Disponível em: <<http://iospress.metapress.com/index/H470T26742N7X506.pdf>>. Citado na página 87.

BENDA, L. E.; POFF, N. L.; TAGUE, C.; PALMER, M. A. How solving. **BioScience**, v. 52, n. 12, 2002. Disponível em: <[http://www.bioone.org/doi/abs/10.1641/0006-3568\(2002\)0525B1127:HTATWW5D2.0.CO3B2](http://www.bioone.org/doi/abs/10.1641/0006-3568(2002)0525B1127:HTATWW5D2.0.CO3B2)>. Citado na página 30.

BERNERS-LEE, T. The world-wide web. **Computer Networks and ISDN Systems**, v. 25, n. 4-5, p. 454–459, 1992. ISSN 01697552. Disponível em: <<http://www.sciencedirect.com/science/article/pii/016975529290039S>>. Citado nas páginas 30 e 77.

_____. **Linked Data - Design Issues**. 2006. Disponível em: <<http://www.w3.org/DesignIssues/LinkedData.html>>. Citado nas páginas 89, 91 e 99.

BERNERS-LEE, T.; CAILLIAU, R.; PELLOW, N.; SECRET, A. The world wide web initiative. **INET International Network Conference**, n. 16 February 2015, 1993. Disponível em: <<https://assets.cs.ncl.ac.uk/seminars/92.pdf>>. Citado nas páginas 30 e 77.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, v. 284, p. 34–43, 2001. ISSN 00368733. Disponível em: <<http://www.nature.com/doi/abs/10.1038/scientificamerican0501-34>>. Citado nas páginas 31 e 77.

BIOCASE. **The BioCASE Access Protocol**. 2013. Disponível em: <<http://biocase.org/products/protocols/index.shtml>>. Citado na página 53.

BIOTA. **Programa Biota-FAPESP Integrando Caracterização, Conservação, Restauração e Uso Sustentável da Biodiversidade**. 2013. Disponível em: <<http://www.biota.org.br/>>. Citado na página 55.

BISBY, F. a. The Quiet Revolution: Biodiversity Informatics and the Internet. **Science**, v. 289, n. 5488, p. 2309–2312, 2000. ISSN 00368075. Disponível em: <<http://www.sciencemag.org/cgi/doi/10.1126/science.289.5488.2309>>. Citado nas páginas 30 e 40.

BORST, W. N. **Construction of Engineering Ontologies for Knowledge Sharing and Reuse**. 243 p. Tese (Doutorado) — Universiteit Twente, 1997. Disponível em: <<http://doc.utwente.nl/17864/>>. Citado na página 31.

BRAHMS. **Botanical Research And Herbarium Management System**. 2013. Disponível em: <<http://herbaria.plants.ox.ac.uk/bol/>>. Citado na página 45.

BREITMAN, K. K. **Web semântica: a internet do futuro**. [S.l.]: LTC, 2006. Citado nas páginas 78 e 82.

CANHOS, D. A. L.; SOUZA, S. de; GIOVANNI, R. de; SIQUEIRA, M. F. de; MARINO, A.; FONSECA, R. L.; CRUZ, B. A.; CANHOS, V. P. **Estudo de Caso Sistemas de Informação On-line: A experiência do CRIA**. [S.l.], 2005. Disponível em: <http://www.cria.org.br/cgee/col/documentos/ntcria{_}vs5.> Citado nas páginas 49, 50, 56 e 57.

CANHOS, V. P. Informática para biodiversidade: padrões, protocolos e ferramentas. **Ciência e Cultura [online]**, v. 55, n. 2, p. 45–47, 2003. ISSN 0009-6725. Disponível em: <http://cienciaecultura.bvs.br/scielo.php?script=sci{_}arttext{\&}pid=S0009-672520030002>. Citado nas páginas 30, 39 e 55.

CANHOS, V. P.; SOUZA, S. de; De Giovanni, R.; CANHOS, D. A. L.; REFERÊNCIA, C. D.; CRIA, A.; TORTIMA, A. R.; GERALDO, B. Global Biodiversity Informatics: setting the scene for a “new world” of ecological forecasting. **Biodiversity Informatics**, v. 1, p. 1–13, 2004. Citado nas páginas 30, 40, 50, 55 e 58.

CARDOSO, S. D. **SWI: Um gazetteer interativo para dados sobre biodiversidade com suporte a web semântica**. 123 p. Tese (master thesis) — Universidade de São Paulo, 2015. Citado na página 125.

CARDOSO, S. D.; AMANQUI, F. K.; SERIQUE, K. J. A.; SANTOS, J. L. C. dos; MOREIRA, D. a. SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data. **Future Generation Computer Systems**, Elsevier B.V., p. –, 2015. ISSN 0167739X. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167739X15001818>>. Citado nas páginas 36, 112, 125, 144 e 145.

CARDOSO, S. D.; SERIQUE, K. J.; AMANQUI, F. K.; SANTOS, J. C. D.; MOREIRA, D. A. A Gazetteer for Biodiversity Data as a Linked Open Data Solution. In: **2014 IEEE 23rd International WETICE Conference**. IEEE, 2014. p. 435–440. ISBN 978-1-4799-4249-7. Disponível em: <<http://ieeexplore.ieee.org/document/6927097/>>. Citado nas páginas 36, 125 e 145.

CBD. **History of the Convention**. 2013. Disponível em: <<http://www.cbd.int/history/>>. Citado nas páginas 59 e 65.

CNPQ. **PELD - Programa de Pesquisa Ecológica de Longa Duração**. 2013. Disponível em: <<http://www.cnpq.br/web/guest/apresentacao7>>. Citado na página 68.

- CORRÊA, P. L. P.; SALVANHA, P.; SARAIVA, A. M.; NETO, P. S.; VALÊNCIO, C. R.; De Souza, R. C. G.; PAULO, S.; JOSÉ, S. A Model for Analysing Data Portal Performance: The Biodiversity Case. **J. Web Eng.**, Rinton Press, Incorporated, Paramus, NJ, v. 12, n. 3-4, p. 232–248, 2013. ISSN 1540-9589. Disponível em: <<http://dl.acm.org/citation.cfm?id=2535629.2535633>>. Citado na página 53.
- COSTELLO, M. J.; APPELTANS, W.; BAILLY, N.; BERENDSOHN, W. G.; JONG, Y. de; EDWARDS, M.; FROESE, R.; HUETTMANN, F.; LOS, W.; MEES, J.; SEGERS, H.; BISBY, F. a. Strategies for the sustainability of online open-access biodiversity databases. **Biological Conservation**, Elsevier Ltd, p. 1–11, nov 2013. ISSN 00063207. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0006320713002747>>. Citado nas páginas 40 e 41.
- COUNCIL, N. R.; OTHERS. **Steps toward large-scale data integration in the sciences: Summary of a workshop**. [S.l.]: National Academies Press, 2010. Citado na página 33.
- CYGANIAK, R.; JENTZSCH, A. **The Linking Open Data cloud diagram**. 2011. Disponível em: <<http://lod-cloud.net/>>. Citado nas páginas 90 e 91.
- DÄÖRING, M.; GIOVANNI, R. D. GBIF Data Access and Database Interoperability: A united protocol for search and retrieval of distributed data. **CRIA-Centro de Referência em Informação Ambiental**, n. September, p. 1–97, 2004. Citado na página 52.
- Darwin Core Task Group, B. I. S. T. **Darwin Core**. 2009. Disponível em: <<http://rs.tdwg.org/dwc/>>. Citado na página 42.
- DEARLE, F. **Groovy for Domain-Specific Languages**. 1st new ed. ed. [S.l.]: Packt Publishing, 2010. ISBN 184719690X, 9781847196903. Citado nas páginas 101 e 102.
- DIGIR. **DiGIR - Table of Contents**. 2013. Disponível em: <<http://digir.sourceforge.net/>>. Citado nas páginas 30 e 51.
- DIMOU, A.; SANDE, M. V.; COLPAERT, P.; VERBORGH, R.; MANNENS, E.; Van De Walle, R. RML: A generic language for integrated RDF mappings of heterogeneous data. **CEUR Workshop Proceedings**, v. 1184, 2014. ISSN 16130073. Citado na página 91.
- DOU. **Política de dados do Programa de Pesquisa em Biodiversidade - PPBio**. 2009. Disponível em: <http://ppbio.inpa.gov.br/sites/default/files/politica{_}dou.> Citado na página 67.
- DUCHARME, B. **Learning Sparql**. [S.l.]: O'Reilly, 2011. 386 p. ISBN 978-1-4493-7146-3. Citado nas páginas 77, 80 e 89.
- EDWARDS, J. L. Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. **Science**, v. 289, n. 5488, p. 2312–2314, sep 2000. ISSN 00368075. Disponível em: <<http://www.sciencemag.org/cgi/doi/10.1126/science.289.5488.2312>>. Citado na página 58.
- _____. Research and societal benefits of the global biodiversity information facility. **BioScience**, v. 54, n. 6, p. 486, 2004. ISSN 0006-3568. Disponível em: <<http://bioscience.oxfordjournals.org/content/54/6/485.full>>. Citado na página 58.
- EGENHOFER, M. J. Spatial SQL: A query and presentation language. **...and Data Engineering, IEEE Transactions on**, v. 6, n. 414, p. 86–95, 1994. Disponível em: <http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=273>. Citado na página 86.

- FEGRAUS, E. H. E.; ANDELMAN, S.; JONES, M. B.; SCHILDHAUER, M. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. **Bulletin of the Ecological Society of America**, Eco Soc America, v. 86, n. 3, p. 158–168, 2005. Disponível em: <[http://www.esajournals.org/doi/pdf/10.1890/0012-9623\(2005\)86\[158:MTVOED\]2.0.CO;2](http://www.esajournals.org/doi/pdf/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2)>. Citado nas páginas 30, 41, 44, 47 e 69.
- FOWLER, M. **Domain-specific languages**. [S.l.]: Pearson Education, 2010. Citado na página 101.
- FREDDY, J.; SU, G. **Ontologias e DSLs na geração de sistemas de apoio à decisão, caso de estudo SustenAgro**. 139 p. Tese (master thesis) — Universidade de São Paulo, 2017. Citado na página 101.
- GADELHA, L. M. R.; GUIMARÃES, P.; MOURA, A. M.; DEBORA, P.; DALCIN, E.; GALL, G.; JR, J. T.; POLTOSI, M.; PORTO, F.; MOURA, F.; LEO, W. V. SiBBr: Uma Infraestrutura para Coleta, Integração e Análise de Dados sobre a Biodiversidade Brasileira. **XXXIV Congresso da Sociedade Brasileira de Computação**, p. 345–352, 2014. Citado na página 59.
- GARDNER, T. a.; BARLOW, J.; CHAZDON, R.; EWERS, R. M.; HARVEY, C. a.; PERES, C. a.; SODHI, N. S. Prospects for tropical forest biodiversity in a human-modified world. **Ecology Letters**, v. 12, n. 6, p. 561–582, jun 2009. ISSN 1461023X. Disponível em: <<http://doi.wiley.com/10.1111/j.1461-0248.2009.01294.x>>. Citado na página 69.
- GBIF. **IPT**. 2013. Disponível em: <<http://www.gbif.org/ipt>>. Citado nas páginas 30 e 54.
- GBIF Secretariat. **GBIF Backbone Taxonomy**. 2017. 1 p. Citado nas páginas 134, 135 e 137.
- GODDARD, A.; WILSON, N.; CRYER, P.; YAMASHITA, G. Data hosting infrastructure for primary biodiversity data. **BMC bioinformatics**, v. 12 Suppl 1, p. S5, jan 2011. ISSN 1471-2105. Disponível em: <<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3287448{&}&tool=pmcentrez{&}rendertype=ab>>. Citado nas páginas 51 e 52.
- GRAU, B. C.; HORROCKS, I.; MOTIK, B.; PARSIA, B.; PATEL-SCHNEIDER, P.; SATTTLER, U. OWL 2: The next step for OWL. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 6, n. 4, p. 309–322, nov 2008. ISSN 15708268. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S1570826808000413>>. Citado na página 83.
- GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, n. 2, p. 199–220, 1993. ISSN 1042-8143. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.7493>>. Citado na página 31.
- GUIZZARDI, G. **Ontological foundations for structural conceptual models**. [S.l.]: CTIT, Centre for Telematics and Information Technology, 2005. Citado nas páginas 95 e 123.
- GÜNTSCH, A.; HYAM, R.; HAGEDORN, G.; CHAGNOUX, S.; RÖPERT, D.; CASINO, A.; DROEGE, G.; GLÖCKLER, F.; GÖDDERZ, K.; GROOM, Q.; HOFFMANN, J.; HOLLEMAN, A.; KEMPA, M.; KOIVULA, H.; MARHOLD, K.; NICOLSON, N.; SMITH, V. S.; TRIEBEL, D. Actionable, long-Term stable and semantic web compatible identifiers for access to biological collection objects. **Database**, v. 2017, n. 1, p. 1–9, 2017. ISSN 17580463. Citado na página 40.
- GURALNICK, R.; HILL, A. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. **Bioinformatics**, v. 25, p. 421–428, 2009. Citado nas páginas 30 e 40.

- HEATH, T.; BIZER, C. Linked data: Evolving the web into a global data space. **Synthesis lectures on the semantic web: theory and technology**, Morgan & Claypool Publishers, v. 1, n. 1, p. 1–136, 2011. ISSN 2160-4711. Disponível em: <<http://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001>>. Citado na página 89.
- HEIDORN, P. B. Biodiversity informatics. **Bulletin of the American Society for Information Science and Technology**, v. 37, n. 6, p. 38–44, aug 2011. ISSN 00954403. Disponível em: <<http://doi.wiley.com/10.1002/bult.2011.1720370612>>. Citado nas páginas 42 e 49.
- HIGGINS, D.; BERKLEY, C.; JONES, M. Managing heterogeneous ecological data using Morpho. In: **Proceedings 14th International Conference on Scientific and Statistical Database Management**. IEEE Comput. Soc, 2002. p. 69–76. ISBN 0-7695-1632-7. ISSN 1099-3371. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1029707>>. Citado nas páginas 47 e 49.
- HOBERN, D.; APOSTOLICO, A.; BELLO, J. C.; CANHOS, D.; DUBOIS, G.; FIELD, D.; HARDISTY, A.; HARRISON, J.; HEIDORN, B.; MATA, E.; PAGE, R.; PARR, C.; PRICE, J.; WILLOUGHBY, S. **Global Biodiversity Informatics Outlook: Delivering Biodiversity Knowledge in the Information Age**. [S.l.], 2013. 44 p. Disponível em: <<http://www.gbif.org/resources/2251>>. Citado nas páginas 30, 40 e 75.
- HOLETSCHEK, J.; DRÖGE, G.; GÜNTSCH, A.; BERENDSOHN, W. G. The ABCD of primary biodiversity data access. **Plant Biosystems**, v. 146, n. 4, p. 771–779, dec 2012. ISSN 1126-3504. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/11263504.2012.740085>>. Citado na página 43.
- HORROCKS, I. Ontologies and the semantic web. **Communications of the ACM**, v. 51, n. 12, p. 58, 2008. ISSN 00010782. Disponível em: <<http://portal.acm.org/citation.cfm?doid=1409360.1409377>>. Citado na página 82.
- ILTER. **The Brazilian LTER Program — ILTER**. 2013. Disponível em: <<http://www.ilternet.edu/member-networks/central-south-america/brazil>>. Citado na página 68.
- JONES, M.; BERKLEY, C.; BOJLOVA, J.; SCHILDHAUER, M. Managing scientific metadata. **IEEE Internet Computing**, v. 5, n. 5, p. 59–68, 2001. ISSN 10897801. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=957896>>. Citado nas páginas 47, 48 e 49.
- KAUPPINEN, T.; ESPINDOLA, G. M. de; JONES, J. Linked brazilian amazon rainforest data. **Semantic Web**, v. 0, n. 0, p. 1–5, 2012. Disponível em: <<http://iospress.metapress.com/index/C788M187215R3390.pdf>>. Citado nas páginas 31 e 121.
- KOENIG, D.; GLOVER, A.; KING, P.; LAFORGE, G.; SKEET, J. **Groovy in action**. 2nd editio. ed. [S.l.]: Manning Publications Co., 2015. Citado na página 102.
- LAPP, H.; MORRIS, R. A.; CATAPANO, T.; HOBERN, D.; MORRISON, N. Organizing our knowledge of biodiversity. **Bulletin of the American Society for Information Science and Technology**, v. 37, p. 38–42, 2011. ISSN 00954403. Citado na página 123.
- LEBO, T.; WILLIAMS, G. T. Converting governmental datasets into linked data. In: **Proceedings of the 6th International Conference on Semantic Systems - I-SEMANTICS '10**. New York, New York, USA: ACM Press, 2010. p. 1. ISBN 9781450300148. Disponível em: <<http://portal.acm.org/citation.cfm?doid=1839707.1839755>>. Citado na página 93.

LUIZAO, F. J.; NOBRE, C. A.; MANZI, A. O. Projeto LBA: Estudando as Complexas Interações da Biosfera com a Atmosfera na Amazônia. **ACTA Amazônica**, v. 35, n. 1-2, p. 109–110, 2005. Citado nas páginas 60 e 61.

MAALI, F.; CYGANIAK, R.; PERISTERAS, V. Re-using Cool URIs: Entity Reconciliation Against LOD Hubs. **LDOW**, 2011. Disponível em: <<http://events.linkedata.org/ldow2011/papers/ldow2011-paper11-maali.pdf>>. Citado na página 93.

MACARTHUR, R. H. **The theory of island biogeography**. [S.l.]: Princeton University Press, 1967. v. 1. Citado na página 69.

MAGNUSSON, W. E.; LIMA, A. P.; LUIZÃO, R.; LUIZÃO, F.; COSTA, F. R. C.; CASTILHO, C. V. de; KINUPP, V. F. RAPELD: a modification of the Gentry method for biodiversity surveys in long-term ecological research sites. **Biota Neotropica**, v. 5, n. 2, p. 19–24, 2005. ISSN 1676-0603. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1676-0603>. Citado na página 67.

MARTINS, S. D. S.; SANDERSON, J. G.; SILVA-JÚNIOR, J. D. S. e. Monitoring mammals in the Caxiuanã National Forest, Brazil – First results from the Tropical Ecology, Assessment and Monitoring (TEAM) program. **Biodiversity and Conservation**, v. 16, n. 4, p. 857–870, 2006. ISSN 0960-3115. Disponível em: <<http://link.springer.com/10.1007/s10531-006-9094-x>>. Citado na página 70.

MAYERNIK, M. S.; BATCHELLER, A. L.; BORGMAN, C. L. How institutional factors influence the creation of scientific metadata. In: **Proceedings of the 2011 iConference on - iConference '11**. New York, New York, USA: ACM Press, 2011. p. 417–425. ISBN 9781450301213. Disponível em: <<http://dl.acm.org/citation.cfm?id=1940761.1940818>>. Citado na página 44.

MEHRING, M.; BALIAN, E. V.; BERHAULT, A.; SCHRAMM, E. **Transdisciplinary research on biodiversity: steps towards integrated biodiversity research**. [S.l.]: European Workshop on “Steps towards Integrated Biodiversity Research”, 2011. Citado na página 29.

MERNIK, M.; HEERING, J.; SLOANE, A. M. When and how to develop domain-specific languages. **ACM Computing Surveys**, v. 37, n. 4, p. 316–344, 2005. ISSN 03600300. Citado na página 101.

MORRIS, R. A. Biodiversity Informatics. In: LEVIN, S. A. (Ed.). **Encyclopedia of Biodiversity (Second Edition)**. Second edi. Waltham: Academic Press, 2013. p. 440–445. ISBN 978-0-12-384720-1. Disponível em: <<http://www.sciencedirect.com/science/article/pii/B9780123847195002628>>. Citado na página 41.

MOURA, A. M. D. C.; PORTO, F.; POLTOSI, M.; PALAZZI, D. C.; MAGALHÃES, P.; VIDAL, V. Integrating Ecological Data Using Linked Data Principles. In: **Joint V Seminar on Ontology Research in Brazil**. [s.n.], 2012. p. 156–167. Disponível em: <http://ceur-ws.org/Vol-938/ontobras-most2012_paper13>. Citado na página 99.

OLIVEIRA, M. L.; BACCARO, F. B.; BRAGA-NETO, R.; MAGNUSSON, W. E. **Reserva Ducke: a biodiversidade amazônica através de uma grade**. Manaus: PPBio, 2008. Citado na página 67.

Open Geospatial Consortium. **OGC GeoSPARQL - A geographic query language for RDF data**. [S.l.], 2012. Disponível em: <<http://www.opengis.net/doc/IS/geosparql/1.0>>. Citado na página 87.

OVASKA, K.; LAAKSO, M.; HAAPA-PAANANEN, S.; LOUHIMO, R.; CHEN, P.; AIT-TOMÄKI, V.; VALO, E.; NÚÑEZ-FONTARNAU, J.; RANTANEN, V.; KARINEN, S.; NOUSIAINEN, K.; MIETTINEN, M.; SAARINEN, L.; KOHONEN, P.; WU, J.; WESTERMARCK, J.; HAUTANIEMI, S.; LAHESMAA-KORPINEN, A.-M.; MIETTINEN, M.; SAARINEN, L.; KOHONEN, P.; WU, J.; WESTERMARCK, J.; HAUTANIEMI, S. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. **Genome Medicine**, v. 2, n. 9, p. 65, 2010. ISSN 1756-994X. Disponível em: <<http://books.google.com/books?hl=en{&}lr={&}id=o3BsnQoA7KoC{&}oi=fnd{&}pg=PT1{&}dq=Steps+Toward+Large-Scale+Data+Integration+in+the+Sciences:+Summary+of+a+Workshop{&}ots=oJzf1JaDyw{&}sig=rBbM7A6GxyCPTNkFrZ6AEpElAn8http://www.ncbi.nlm.nih.gov/pubmed>>. Citado na página 33.

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. 345–54, sep 2008. ISSN 1477-4054. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/18445641>>. Citado na página 41.

PEIXOTO, A. L.; BARBOSA, M. R.; MENEZES, M.; MAIA, L. C. Diretrizes e estratégias para a modernização de coleções biológicas brasileiras e a consolidação de sistemas integrados de informação sobre biodiversidade. **Brasília: Centro de Gestão e Estudos Estratégicos: Ministério da Ciência e Tecnologia**, Brasília, p. 145–182, 2006. Citado na página 52.

PEIXOTO, A. L.; BARBOSA, M. R. D. V.; CANHOS, D. A. L.; MAIA, L. C. Coleções botânicas: objetos e dados para a ciência. **Granato, M. & Rangel, M. (Org.). Cultura material e patrimônio de Ciência e Tecnologia**, p. 315–326, 2009. Citado na página 66.

PENNINGTON, D. D. Supporting large-scale science with workflows. **Proceedings of the 2nd workshop on Workflows in support of large-scale science - WORKS '07**, ACM Press, New York, New York, USA, p. 45, 2007. Disponível em: <<http://dl.acm.org/citation.cfm?doid=1273360.1273369>>. Citado nas páginas 29 e 33.

PENNINGTON, D. D.; SIMPSON, G. L.; MCCONNELL, M. S.; FAIR, J. M.; BAKER, R. J.; CONNELL, M. S. M. Transdisciplinary Research, Transformative Learning, and Transformative Science. **BioScience**, v. 63, n. 7, p. 564–573, jul 2013. ISSN 00063568. Disponível em: <<http://www.jstor.org/stable/info/10.1525/bio.2013.63.7.9>>. Citado na página 29.

PERES, C. a.; GARDNER, T. a.; BARLOW, J.; ZUANON, J.; MICHALSKI, F.; LEES, A. C.; VIEIRA, I. C. G.; MOREIRA, F. M. S.; FEELEY, K. J. Biodiversity conservation in human-modified Amazonian forest landscapes. **Biological Conservation**, Elsevier Ltd, v. 143, n. 10, p. 2314–2327, oct 2010. ISSN 00063207. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0006320710000224>>. Citado na página 69.

PNB. **Política Nacional da Biodiversidade**. Brasília: [s.n.], 2002. Disponível em: <<http://www.mma.gov.br/biodiversidade/biodiversidade-brasileira/politica-nacional-da-biodiver>>. Citado nas páginas 65 e 66.

POLLOCK, J. T. **Semantic Web for dummies**. [S.l.]: John Wiley & Sons, 2009. Citado nas páginas 78, 79 e 81.

Portal Brasil. **CNPq recebe pesquisadores que participam da reunião do PELD** — Portal Brasil. 2013. Disponível em: <<http://www.brasil.gov.br/educacao/2013/09/cnpq-recebe-pesquisadores-que-participam-da-reuniao-do-peld>>. Citado na página 68.

PPBIO. Núcleo Regional BioGeoInformática | PPBio Amazônia Oriental. 2013. Disponível em: <<http://ppbio.museu-goeldi.br/?q=pt-br/n{ú}cleo-de-biogeoinfor>>. Citado na página 67.

Programa LBA. **Sobre o Programa LBA**. 2007. Disponível em: <<http://lba.inpa.gov.br/lba/?p=intro{&}>>>. Citado nas páginas 61 e 65.

PYLRO, V. S.; ROESCH, L. F. W.; ORTEGA, J. M.; AMARAL, A. M. do; TÓTOLA, M. R.; HIRSCH, P. R.; ROSADO, A. S.; GÓES-NETO, A.; da Costa da Silva, A. L.; ROSA, C. A.; MORAIS, D. K.; ANDREOTE, F. D.; DUARTE, G. F.; MELO, I. S. de; SELDIN, L.; LAMBAIS, M. R.; HUNGRIA, M.; PEIXOTO, R. S.; KRUGER, R. H.; TSAI, S. M.; AZEVEDO, V. Brazilian Microbiome Project: Revealing the Unexplored Microbial Diversity-Challenges and Prospects. **Microbial Ecology**, p. 4–8, nov 2013. ISSN 1432-184X. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/24173537>>. Citado na página 58.

RANDALL, D. D.; ABBOTT, M. R.; BRUER, J. T.; DROEGEMEIER, K.; HASTINGS, D. E.; HESS, K.; LANZEROTTI, L. J.; LESHNER, A. I. Enhancing Support of Transformative Research at the National Science Foundation. **Task Force on Transformative Research, NSB-07-32, National Science Foundation**, 2007. Citado na página 29.

REICHMAN, O. J.; JONES, M. B.; SCHILDHAUER, M. P. Challenges and opportunities of open data in ecology. **Science (New York, N.Y.)**, v. 331, n. 6018, p. 703–5, feb 2011. ISSN 1095-9203. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/21311007>>. Citado na página 79.

ROBERTSON, T. **Integrated Publishing Toolkit (IPT)**. 2011. Disponível em: <<http://www.gbif.org/ipt>>. Citado nas páginas 30 e 54.

ROBERTSON, T.; DÖRING, M.; GURALNICK, R.; BLOOM, D.; WIECZOREK, J.; BRAAK, K.; OTEGUI, J.; RUSSELL, L.; DESMET, P. The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. **PLoS ONE**, v. 9, n. 8, p. e102623, aug 2014. ISSN 1932-6203. Disponível em: <<http://dx.plos.org/10.1371/journal.pone.0102623>>. Citado nas páginas 42, 53 e 54.

San Gil, I.; BAKER, K.; CAMPBELL, J.; DENNY, E. G.; VANDERBILT, K.; RIORDAN, B.; KOSKELA, R.; DOWNING, J.; GRABNER, S.; MELENDEZ, E.; OTHERS; GIL, I. S. The Long-Term Ecological Research community metadata standardisation project: a progress report. **International Journal of Metadata, Semantics and Ontologies**, Inderscience, v. 4, n. 3, p. 141–153, 2009. Disponível em: <<http://inderscience.metapress.com/index/D0231885742P87T5.pdf>>. Citado na página 44.

SANTOS, J. L. dos; Campos dos Santos, J. L. **A biodiversity information system in an open data/metadatabase architecture**. Tese (Doutorado) — ITC, Enschede, 2003. Disponível em: <<http://doc.utwente.nl/41436/>>. Citado nas páginas 41 e 62.

SANTOS, J. L. dos; de Magalhães Netto, J. F.; CASTRO, A. N. de; ALBUQUERQUE, A. C. F.; FERNEDA, E.; ALONSO, L.; ROCHA, R. L. d. C.; MENDONÇA, D. T. de; NETTO, J. F. d. M.; Campos dos Santos, J. L.; de Magalhães Netto, J. F.; CASTRO, A. N. de; ALBUQUERQUE, A. C. F.; FERNEDA, E.; ALONSO, L.; ROCHA, R. L. d. C.; MENDONÇA, D. T. de; NETTO, J. F. d. M. Ontologias para Interoperabilidade de Modelos e Sistemas de Informação de Biodiversidade. **ceur-ws.org**, Gramado, Brazil, v. 728, p. 6, may 2011. Disponível em: <<http://ceur-ws.org/Vol-728/paper8.pdf>>. Citado nas páginas 39, 68, 95, 97 e 123.

SARKAR, I. N. Biodiversity informatics: organizing and linking information across the spectrum of life. **Briefings in bioinformatics**, v. 8, n. 5, p. 347–57, sep 2007. ISSN 1477-4054. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17704120>>. Citado na página 39.

SCHARFFE, F.; ATEMEZING, G.; TRONCY, R. Enabling linkeddata publication with the datalift platform. **Proc. AAAI workshop on ...**, 2012. Disponível em: <<http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/viewFile/5349/5678>>. Citado na página 92.

Secretariat of the Convention on Biological Diversity. **Handbook of the Convention on Biological Diversity**. 3rd editio. ed. Earthscan, 2005. v. 114. 467 p. ISSN 00063207. ISBN 9292250116. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0006320703000533>>. Citado na página 59.

SEGARAN, T.; EVANS, C.; TAYLOR, J. **Programming the semantic web**. [S.l.]: O'Reilly Media, 2009. ISBN 978-0-596-15381-6. Citado na página 85.

SERIQUE, K. J. A.; SANTOS, J. L. C. dos; ALBUQUERQUE, A. C. F. Mo Porã: Uma ferramenta para o gerenciamento distribuído de repositórios de dados científicos na web. In: EMILIO, T.; LUISÃO, F. (Ed.). **Cenários para a Amazônia: clima, biodiversidade e uso da terra**. Manaus: Editora INPA, 2014. cap. 13, p. 194. ISBN 978-85-211-0126-0. Citado nas páginas 36, 128, 143 e 145.

SERIQUE, K. J. d. A.; SANTOS, J. L.; MOREIRA, D. d. A.; OTHERS. BioDSL: a domain-specific language for mapping and dissemination of biodiversity data in the LOD. **Congresso da Sociedade Brasileira de Computação, XXXVI; Brazilian e-Science Workshop, X**, p. 285–293, 2016. Citado nas páginas 36, 144 e 145.

SETTE, L. D.; PAGNOCCA, F. C.; RODRIGUES, A. Microbial culture collections as pillars for promoting fungal diversity, conservation and exploitation. **Fungal genetics and biology : FG & B**, Elsevier Inc., v. 60, p. 2–8, 2013. ISSN 1096-0937. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/23872281>>. Citado na página 59.

SEYED, P.; CHASTAIN, K.; ASHBY, B.; LIU, Y.; LEBO, T.; PATTON, E.; MCGUINNESS, D. Semanteco annotator. **CEUR Workshop Proceedings**, v. 1035, p. 161–164, 2013. ISSN 16130073. Disponível em: <http://iswc2013.semanticweb.org/sites/default/files/iswc{_}demo{_}>>. Citado na página 93.

SHADBOLT, N.; BERNERS-LEE, T.; HALL, W. The Semantic Web Revisited. **IEEE Intelligent Systems**, v. 21, n. 3, p. 96–101, jan 2006. ISSN 1541-1672. Citado nas páginas 31 e 91.

SHANMUGHAVEL, P. An overview on biodiversity information in databases. **Bioinformation**, p. 367–369, 2007. Disponível em: <<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1891718/>>. Citado na página 121.

SIBBR. **SiBBr**. 2013. Disponível em: <<http://www.sibbr.gov.br/>>. Citado na página 59.

SINBIOTA. **SinBiota**. 2013. Disponível em: <<http://sinbiota.biota.org.br/>>. Citado na página 55.

SPECIESLINK. **speciesLink: Sistema de Informação Distribuído para Coleções Biológicas**. 2013. Disponível em: <<http://splink.cria.org.br/>>. Citado na página 55.

SPECIFY. **Welcome to Specify 6 | Specify 6**. 2013. Disponível em: <<http://specifysoftware.org/>>. Citado na página 46.

STADLER, C.; UNBEHAUEN, J.; WESTPHAL, P.; SHERIF, M. A.; LEHMANN, J. Simplified RDB2RDF Mapping. **Proceedings of the 8th Workshop on Linked Data on the Web (LDOW2015), Florence, Italy**, 2015. Disponível em: <svn.aksw.org/papers/2015/LDOW/{_}SML/paper-camery-ready/{_}publ>. Citado nas páginas 92 e 93.

TDWG. **TDWG: DarwinCore**. 2009. Disponível em: <<http://www.tdwg.org/activities/darwincore/>>. Citado na página 44.

_____. **TAPIR - TDWG Access Protocol for Information Retrieval**. 2010. Disponível em: <<http://www.tdwg.org/dav/subgroups/tapir/1.0/docs/>>. Citado nas páginas 30 e 52.

TEAM. **About TEAM**. 2016. Disponível em: <<http://www.teamnetwork.org/about>>. Citado na página 70.

UNBEHAUEN, J.; HELLMANN, S.; AUER, S.; STADLER, C. Knowledge extraction from structured sources. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 7538, p. 34–52, 2012. ISSN 03029743. Citado na página 92.

Van Deursen, A.; KLINT, P. Domain-Specific Language Design Requires Feature Descriptions. **Journal of Computing and Information Technology**, v. 10, n. 1, p. 1–17, 2002. ISSN 1330-1136. Disponível em: <<http://cit.srce.unizg.hr/index.php/CIT/article/view/1465>>. Citado na página 101.

W3C. **OWL Web Ontology Language Overview**. 2004. Disponível em: <<http://www.w3.org/TR/owl-features/>>. Citado na página 83.

_____. **RDF Primer**. 2004. Disponível em: <<http://www.w3.org/TR/rdf-primer/>>. Citado nas páginas 32 e 80.

_____. **SPARQL Query Language for RDF**. 2008. Disponível em: <<http://www.w3.org/TR/rdf-sparql-query/>>. Citado na página 85.

_____. **OWL 2 Web Ontology Language Document Overview (Second Edition)**. 2012. Disponível em: <<http://www.w3.org/TR/owl2-overview/>>. Citado na página 83.

_____. **Semantic Web - W3C**. 2013. Disponível em: <<http://www.w3.org/standards/semanticweb/>>. Citado na página 78.

_____. **SPARQL 1.1 Overview**. 2013. Disponível em: <<http://www.w3.org/TR/sparql11-overview/>>. Citado na página 85.

_____. **SweoIG/TaskForces/CommunityProjects/LinkingOpenData - W3C Wiki**. 2013. Disponível em: <<http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>>. Citado na página 90.

WAAL, S. van der; WECEL, K.; ERMILOV, I.; JANEV, V.; MILOŠEVIĆ, U.; WAINWRIGHT, M. Lifting Open Data Portals to the Data Web. In: **Linked Open Data—Creating Knowledge Out of Interlinked Data**. Springer, 2014. v. 8661, p. 175–195. Disponível em: <<http://link.springer.com/10.1007/978-3-319-09846-39>>. Citado na página 99.

WALLS, R. L.; DECK, J.; GURALNICK, R.; BASKAUF, S.; BEAMAN, R.; BLUM, S.; BOWERS, S.; BUTTIGIEG, P. L.; DAVIES, N.; ENDRESEN, D.; GANDOLFO, M. A.; HANNER, R.; JANNING, A.; KRISHTALKA, L.; MATSUNAGA, A.; MIDFORD, P.; MORRISON, N.; TUAMA, É. Ó.; SCHILDHAUER, M.; SMITH, B.; STUCKY, B. J.; THOMER, A.; WIECZOREK, J.; WHITACRE, J.; WOOLEY, J. Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. **PLoS ONE**, v. 9, n. 3, mar 2014. ISSN 1932-6203. Disponível em: <<http://dx.plos.org/10.1371/journal.pone.0089606>>. Citado na página 122.

WIECZOREK, J. R.; BLOOM, D.; GURALNICK, R. P.; BLUM, S.; DÖRING, M.; GIOVANNI, R.; ROBERTSON, T.; VIEGLAIS, D. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. **PLoS ONE**, v. 7, n. 1, p. 8, jan 2012. ISSN 1932-6203. Disponível em: <<http://dx.plos.org/10.1371/journal.pone.0029715>>. Citado nas páginas 30, 33, 42, 43 e 103.

YESSON, C.; BREWER, P. W.; SUTTON, T.; CAITHNESS, N.; PAHWA, J. S.; BURGESS, M.; GRAY, W. A.; WHITE, R. J.; JONES, A. C.; BISBY, F. A.; CULHAM, A. How global is the global biodiversity information facility? **PloS one**, v. 2, n. 11, p. e1124, 2007. Citado nas páginas 58 e 125.

ZUQUIM, G.; COSTA, F. R. C.; PRADO, J. Redução de esforço amostral vs. retenção de informação em inventários de pteridófitas na Amazônia Central. **Biota Neotropica**, v. 7, n. 3, p. 217–223, 2007. Disponível em: <<http://www.scielo.br/pdf/bn/v7n3/22.pdf>>. Citado na página 67.