Pontifícia Universidade Católica
do Rio de Janeiro

Ricardo Luís da Costa Rocha

Web Semântica Aplicada às Coleções Biológicas do INPA

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio.

Orientador: Prof. Daniel Schwabe

Rio de Janeiro
Julho de 2012

Ricardo Luís da Costa Rocha

Web Semântica Aplicada às Coleções Biológicas do INPA

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Daniel Schwabe
Orientador
Departamento de Informática - PUC-Rio

Prof. Hermann Haeusler
Departamento de Informática - PUC-Rio

Profª. Maria Luiza Machado Campos
UFRJ

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico - PUC-Rio

Rio de Janeiro, 30 de julho de 2012

Ricardo Luís da Costa Rocha

Graduou-se em Processamento de Dados pela Universidade Federal do Amazonas em 2000. Especializou-se em Tecnologia Web também pela Universidade Federal do Amazonas em 2007. Atua como tecnologista pleno no Instituto Nacional de Pesquisas da Amazônia. Tem experiência na área de Ciência da Computação, com ênfase em Redes de Computadores, atuando principalmente nos seguintes temas: gerência de redes de computadores, serviços de redes em software livre, segurança em redes, tecnologia Web.

Ficha Catalográfica

Rocha, Ricardo Luís da Costa

Web semântica aplicada às coleções biológicas do INPA / Ricardo Luís da Costa Rocha ; orientador: Daniel Schwabe. – 2012.
120 f. : il. (color.) ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2012.
Inclui bibliografia

1. Informática – Teses. 2. Web semântica. 3. Coleções biológicas. 4. Gestão do conhecimento. I. Schwabe, Daniel. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Para meus pais pelo amor incondicional, apoio e confiança.

Agradecimentos

À Deus pelo sopro da vida.

Ao meu orientador Professor Daniel Schwabe por acreditar, pela paciência, estímulo e parceria para realização deste trabalho.

Ao meu professor e amigo José Laurindo Campos dos Santos, pela co-orientação, pelas importantes contribuições, ensinamentos e palavras de apoio.

Ao INPA, à FAPEAM e à PUC-Rio pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Aos meus pais, pela educação, atenção e carinho de todas as horas.

Às minhas amigas Andréa Albuquerque, Maysa Roriz, Regina Marques e Rosana Gemaque por todo apoio, colaboração e compreensão.

Aos meus amigos José Eduardo, Percy Henrique, Renato Javier, Ricardo Henriques e Ricardo Rios pelo companheirismo durante a jornada do mestrado.

À Profa. Maria Luiza e seu grupo de pesquisa, principalmente Fabrício Firmino e João Vitor, por me repassarem o conhecimento técnico necessário para a realização de algumas fases da dissertação.

À Profa. Maria Luiza, ao Prof. Hermann e ao Prof. Casanova por atenderam prontamente minha solicitação e participarem da Comissão Examinadora deste trabalho.

A todos os amigos, principalmente a D. Mundinha, a D. Deuza e o Sanderley, e familiares que de uma forma ou de outra me ajudaram ou me incentivaram.

Especialmente, à Colleen, companheira que tive a graça de encontrar em momento tão desafiador da minha vida, pelo carinho, paciência, compreensão e por apostar em mim todos os dias.

Resumo

Rocha, Ricardo Luís da Costa; Schwabe, Daniel. **Web Semântica Aplicada às Coleções Biológicas do INPA.** Rio de Janeiro, 2012. 120p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A Web Semântica permite a divulgação de dados na Internet através de um formato comum com o objetivo de integrar ou combinar bases de dados provenientes de diversas fontes. O Instituto Nacional de Pesquisas da Amazônia - INPA possui várias coleções de dados, principalmente científicos, que podem ser divulgadas e utilizadas na pesquisa e desenvolvimento da Amazônia e para o progresso da ciência. O objetivo deste trabalho é investigar como a utilização das tecnologias da Web Semântica, dentre elas os recursos da ferramenta Rexplorator, pode melhorar o processo de pesquisa, através do processamento da semântica, das coleções de dados biológicos do instituto. A abordagem utilizada é de desenvolver casos de uso junto com os próprios pesquisadores, através de operações simples em cima dos modelos RDFS (Resource Description Framework Schema) das próprias bases. Os casos de uso poderão ser reutilizados por outros pesquisadores, inclusive de domínios de pesquisa diferentes. Neste processo de reutilização é possível que os casos de uso sejam customizados e evoluídos colaborativamente no próprio ambiente em que foram desenvolvidos. Como resultado do processo são geradas aplicações Web que abstraem os modelos RDF (Resource Description Framework) nos quais os dados estão representados tornando possível o acesso às informações por outros pesquisadores que não conhecem esses modelos. Essa facilidade de acesso, além de permitir consultas a bases semânticas por usuários leigos em um dado domínio de pesquisa, também visa permitir que pesquisadores possam realizar consultas transdisciplinares enriquecendo sua visão no desenvolvimento da pesquisa, bem como seu poder nas tomadas de decisões políticas, econômicas e sociais, e, consequentemente, uma melhor gestão do conhecimento.

Palavras-chave

Web Semântica; coleções biológicas; gestão do conhecimento.

Abstract

Rocha, Ricardo Luís da Costa; Schwabe, Daniel (Advisor). **Semantic Web Applied to INPA's Biological Collections**. Rio de Janeiro, 2012. 120p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The Semantic Web enables the dissemination of data on the Internet through a common format in order to integrate or combine databases from different sources. The National Institute for Amazonian Research (INPA) has several collections of data, mainly scientific, that can be disclosed and used in the research and development of the Amazon as well as for the advancement of science. The objective of this study is to investigate how the use of Semantic Web technologies, among them the tool Rexplorator, can improve the research process by processing the semantics in collections of biological data. Use cases are developed with input from INPA researchers. Queries are constructed based on RDFS (Resource Description Framework Schema) created for INPA's existing collections. Use cases can be reused by other researchers, including researchers from different fields. In this process of reuse, the customization and collaborative development of use cases is possible. The result of this process is the generation of Web applications that abstract the RDF model on which data are represented. Consequently, other researchers unfamiliar with the RDF model are also able to access information. In addition to enabling semantic queries in databases by lay users in a given field of research, this ease of access enables researchers to make transdisciplinary queries enriching their vision of research development, as well as their power in political, economic and social decision-making, and hence better knowledge management.

Keywords

Semantic Web; biological collections; knowledge management.

Sumário

1 Introdução 18

1.1. Contextualização 18

1.2. Cenário da pesquisa no INPA 22

1.3. Motivação do trabalho 23

1.4. Objetivos 24

1.5. Organização da dissertação 26

2 Fundamentos 28

2.1. Tecnologias: da Web Estática à Web Semântica 28

2.1.1. Web Semântica 31

2.2. A evolução da representação de dados biológicos 39

2.2.1. *Darwin Core* 40

2.2.2. *Access to Biological Collection Data* (ABCD) 41

2.2.3. *Plinian Core* 42

2.2.4. *Ecological Metadata Language* (EML) 42

2.2.5. *Clustered Object Schema for INPA's Biodiversity Data Collections* (CLOSi) 43

2.3. Evolução de sistemas para gestão de dados e informações de biodiversidade em ambiente integrado 43

2.3.1. SIBs e suas aplicações para ambiente integrado 46

2.4. Paralelo entre evolução Web e evolução das coleções biológicas 48

3 Metodologia 50

3.1. Introdução 50

3.2. Criação das bases de dados semânticas 50

3.2.1. Esquema de dados e ontologia de domínio: pontos de partida 51

3.2.2. Etapas do processo 55

3.2.3. Ferramentas selecionadas 57

3.2.4. Trabalhos relacionados 59

3.3. Construção das consultas 60

3.3.1. O Rexplorator 60

3.3.2. Trabalhos relacionados e outras ferramentas 64

4 Exemplo de Aplicação 66

4.1. Introdução 66

4.2. Registrando espécimes sem semântica 68

4.3. Criação da base de dados semântica ligada 70

4.3.1. Coleta de conjuntos de dados biológicos 70

4.3.2. Modelagem de dados 71

4.3.3. Seleção e extração de dados sobre aves 74

4.3.4. Curagem dos dados digitais 75

4.3.5. Mapeamento para base de dados triplificada 77

4.3.6. Ligação da base de dados triplificada com a LOD 79

4.3.7. Armazenamento da base de dados triplificada 81

4.4. Exploração da base RDF com o Rexplorator 82

4.4.1. Casos de uso 82

4.4.2. Implementação da aplicação Web 84

4.4.3. Telas da aplicação implementada para consultas à base de dados 98

4.5. Impacto nos processos de pesquisa 110

5 Considerações Finais 112

5.1. Resultados alcançados 113

5.2. Contribuições do trabalho 113

5.3. Trabalhos futuros 115

6 Referências Bibliográficas 116

Lista de figuras

Figura 1 - Linha do tempo para a evolução da Web 29

Figura 2 - As ligações na Web de Documentos e na Web de Dados 31

Figura 3 - Exemplo de documentos da coleta da classe Arachnida UFAM/INPA 45

Figura 4 - Evolução do cenário das fontes de dados e informações sobre a biodiversidade amazônica 49

Figura 5 - Clusters e estrutura dos relacionamentos do esquema CLOSi 53

Figura 6 - Visão Geral da OntoBio 55

Figura 7 - Modelo do Rexplorator 61

Figura 8 - Itens derivados de um espécime 69

Figura 9 - Modelo ER para a Coleção de Aves da Amazônia do INPA 72

Figura 10 - Modelo lógico do banco de dados da Coleção de Aves da Amazônia do INPA 73

Figura 11 - Exemplo de erro de registro de dados 76

Figura 12 - Identificação, edição e correção de erro utilizando o Google Refine 76

Figura 13 - *Amazonian Birds Collection Ontology* (ABC) 78

Figura 14 - Após a criação do conjunto de triplas “entidades bióticas” que possui a propriedade “eRepresentadaPor” na posição de predicado 84

Figura 15 - Criação do conjunto “espécimes por entidade biótica” 85

Figura 16 - Entidade biótica dos espécimes sendo parametrizada para tornar possível a reutilização da consulta 86

Figura 17 - Definição do conjunto de triplas como ponto de partida para o caso de uso 87

Figura 18 - Adicionando o comportamento de passagem de valor para parâmetro ao conjunto de triplas 88

Figura 19 - Indicação que o valor do recurso na posição do sujeito que for selecionado será usado como parâmetro do outro conjunto de triplas 88

Figura 20 - Criação de um conjunto com as propriedades desejadas através de uma operação SPO 89

Figura 21 - Parametrizando o conjunto de triplas criado e renomeado 90

Figura 22 - Conjuntos de triplas resultantes que implementam o primeiro caso de uso 90

Figura 23 - Conjuntos de triplas resultantes que implementam o segundo caso de uso com o conceito de consulta subjacente 91

Figura 24 - Criação de um conjunto que possui os táxons na posição de sujeito e os nomes científicos na posição de objeto 93

Figura 25 - Opção do menu principal para a criação de uma entrada de texto e transdutor renomeado com a semântica adequada 93

Figura 26 - Conjunto “Classificação Taxonômica” gerado a partir do conjunto “Nomes Científicos” e do transdutor de texto “Consulta Nome Científico” 94

Figura 27 - Conjunto de triplas gerado pela palavra-chave 95

Figura 28 - Conjunto com informações sobre o nome científico pesquisado 96

Figura 29 - Inclusão de repositórios da remotos para a geração de consultas 97

Figura 30 - Adição de informações da LOD através de links de identidade para conjuntos de dados externos 98

Figura 31 - Tela inicial da aplicação de consulta à Coleção de Aves do INPA 98

Figura 32 - (1) Seleção da consulta de espécimes por entidade biótica; (2) Seleção da entidade biótica da lista de entidades disponíveis 99

Figura 33 - (3) Espécimes por entidade biótica; (4) Informações adicionais sobre o espécime 100

Figura 34 - (1) Seleção da consulta de espécimes por táxon; (2) Seleção do táxon da lista de categorias taxonômicas disponíveis; (3) Seleção da entidade biótica 101

Figura 35 - (4) Espécimes por entidade biótica; (5) Informações adicionais sobre o espécime 101

Figura 36 - (1) Seleção da consulta de espécimes por coleção; (2) Seleção da Coleção de Peles da lista de coleções disponíveis 102

Figura 37 - (3) Espécimes por coleção; (4) Informações adicionais sobre o espécime 102

Figura 38 - (1) Seleção da consulta de espécimes por preparador; (2) Seleção do preparador da lista de preparadores disponíveis 103

Figura 39 - (3) Espécimes por preparador; (4) Informações adicionais sobre o espécime 104

Figura 40 - (1) Seleção da consulta de espécimes por coleta; (2) Seleção de coleta da lista de coletas disponíveis. 104

Figura 41 - (3) Espécimes por coleta; (4) Informações adicionais sobre o espécime 105

Figura 42 - (1) Seleção da consulta de espécimes por localização; (2) Seleção do local da lista de localizações disponíveis; (3) Seleção de coleta realizada na localização 106

Figura 43 - (4) Espécimes por localização; (5) Informações adicionais sobre o espécime 106

Figura 44 - (1) Seleção da consulta por nome científico, (2) Busca por nome científico e (3) Tripla resultante da consulta por nome científico 107

Figura 45 - (4) Espécimes por entidade biótica; (5) Informações sobre o espécime 108

Figura 46 - (6) Lista de informações da base local sobre a espécie e seleção de recurso da base TaxonConcept; (7) Informações TaxonConcept sobre a espécie 109

Lista de quadros

Quadro 1 - Serialização RDF/XML 35

Quadro 2 - Serialização RDFa 36

Quadro 3 - Serialização Turtle 36

Quadro 4 - Serialização com N-Triples 37

Quadro 5 - Serialização N-Triples para *Harpia harpyja* 78

Quadro 6 - Exemplos de propriedades de objetos 79

Quadro 7 - Ligação de recursos para *Harpia harpyja* 80

Quadro 8 - Código do Operador de filtro por palavra-chave 92

Lista de tabelas

Tabela 1 - Resumo da curagem dos dados digitais da Coleção de Aves do INPA 77

Lista de Abreviaturas e Siglas

ABC	*Amazonian Birds Collection*
ABCD	*Access to Biological Collection Data*
ADAPTA	INCT de Adaptações da Biota Aquática da Amazônia
API	*Application Programming Interfaces*
ASC	*Association of Systematic Collections*
ASKAP	*Australian Square Kilometre Array Pathfinder*
BBC	*British Broadcasting Corporation*
CENBAM	INCT de Estudos Integrados da Biodiversidade Amazônica
CLOSi	*Clustered Object Schema for INPA's Biodiversity Data Collections*
CRBio	Sistema Costarricense de Información sobre Biodiversidad
CBRO	Comitê Brasileiro de Registros Ornitológicos
CSV	*Comma-Separated Values*
CTPETRO	Rede temática cooperativa entre instituições de ensino superior e de ciência e tecnologia na Amazônia
DCMI	*Dublin Core Metadata Iniciative*
DERI	*Digital Enterprise Research Institute*
DOM	*Document Object Model*
EEA	*European Environment Agency*
EFG	*Electronic Field Guide*
EIONET	*European Environment Information and Observation Network*
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
EML	*Ecological Metadata Language*
ER	*Entity Relationship Model*
ETL	*Extraction, Transformation, Loading*
GBIF	*Global Biodiversity Information Facility*
GBIF.ES	*Global Biodiversity Information Facility in Spain*
GEOMA	Rede Temática de Pesquisa em Modelagem Ambiental da Amazônia
HTML	Hyper*Text Markup Language*
HTTP	*HyperText Transfer Protocol*
IA	Inteligência Artificial
IEPA	Instituto de Pesquisas Científicas e Tecnológicas do Estado do Amapá
INCT	Instituto Nacional de Ciência e Tecnologia
INPA	Instituto Nacional de Pesquisas da Amazônia
IUCN	*International Union Conservation of Nature*
ITIS	*Integrated Taxonomic Information System*
LBA	Programa de Grande Escala da Biosfera-Atmosfera da Amazônia
LIS	Laboratório de Interoperabilidade Semântica
LHC	*Large Hadrons Colider*
LOD	*Linked Open Data*
MADEIRAS	INCT de Madeiras da Amazônia
MPEG	Museu Paraense Emílio Goeldi
MVC	*Model-View-Controller*

NCEAS	*National Center for Ecological Analysis and Synthesis*
OPM	*Object-Protocol Model*
OWL	*Ontology Web Language*
PCAC	Programa de Coleções e Acervos Científicos
PGV	*Paged Graph Visualization*
PPBio	Programa de Pesquisa em Biodiversidade
RDF	*Resource Description Framework*
RDFa	*Resource Description Framework in Attributes*
RDFS	RDF Schema
SBC	Sociedade Brasileira de Computação
SERVAMB	INCT dos Serviços Ambientais da Amazônia
SHDM	Semantic Hypermedia Design Method
SIB	Sistemas de Informação de Biodiversidade
Silvolab	Laboratório de Silviculturana na Guiana Francesa
SPARQL	*SPARQL Protocol and RDF Query Language*
SWRL	*Semantic Web Rule Language*
TDWG	*Biodiversity Information Standards* - também conhecido como *Taxonomic Databases Working Group*
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
URN	*Uniform Resource Name*
W3C	*World Wide Web Consortium*
WS	Web Semântica
WWF	*World Wild Fund*
WWW	*World Wide Web*
XML	*eXtensible Markup Language*

1
Introdução

1.1.
Contextualização

> A ciência está se tornando cada vez mais dependente de dados, ainda que as tecnologias de dados tradicionais não tenham sido projetadas para a escala e heterogeneidade dos dados no mundo moderno. Projetos como o *Large Hadrons Colider* (LHC) e o *Australian Square Kilometre Array* Pathfinder (ASKAP) irão gerar *petabytes* de dados que devem ser analisados por cientistas trabalhando em vários países e falando muitos idiomas diferentes. A facilitação digital ou eletrônica da ciência, ou *e-Science*, é agora essencial e está se tornando generalizada.
>
> [Fox & Hendler, 2009]

Instituições de pesquisas no Brasil, no contexto de suas atividades científicas sobre a biodiversidade, têm coletado dados oriundos de experimentos, expedições, inventários, etc., para as devidas análises e processamento, objetivando a produção do tão desejado conhecimento em diferentes domínios. Ao longo do século XX a ineficiência na custódia e gestão dos dados tornou a localização e acesso aos dados um desafio gigantesco. Com o advento da facilitação de acesso via Web, repositórios para grande volume de dados, baixo custo na comunicação, e ferramentas para exploração disponíveis, estão levando os dados a condição de *commodities* sendo, inegavelmente, essenciais para a pesquisa científica, uma vez que poderão ser amplamente utilizados e reutilizados.

Uma lista de aplicações consideradas importantes no contexto político social, por exemplo, avaliação de impacto ambiental, definição de áreas de preservação ambiental, proteção de espécies ameaçadas, recuperação de áreas degradadas, bioprospecção, estabelecimento de políticas públicas, legislação ambiental, etc., são completamente dependentes de dados e, importante enfatizar, dados de qualidade e preferencialmente certificados [Campos dos Santos et al., 2000; Umminger & Young, 1997; Albuquerque, 2011].

O delineamento de um experimento, a coleta de dados, a curagem, a análise, a visualização dos dados e a comunicação de resultados são as fases do ciclo de vida da pesquisa.

A infraestrutura cibernética, provida atualmente pela e-Ciência ou *e-Science*, possibilita aos cientistas desenvolver pesquisa de maneira mais rápida, mais eficiente ou até mesmo diferente de práticas comuns. e-Ciência é, em resumo, uma ferramenta que provê aos pesquisadores uma estrutura para armazenar, interpretar, analisar e disponibilizar dados em rede para outros grupos de pesquisas. As previsões de analistas indicam que e-Ciência irá revolucionar os mais diversos mecanismos da pesquisa científica, iniciando com a pesquisa teórica básica, testes através de simulações, testes controlados, coleção de dados de forma organizada e na interpretação dos dados.

A alta capacidade de processamento disponível e a possibilidade de integração de grandes volumes de dados distribuídos em diferentes fontes é uma realidade e viabiliza projetos colaborativos inter, multi ou transdisciplinares.

Paralelamente ou concomitantemente, ocorreu o estabelecimento de uma ciência de estudo da evolução da Rede Mundial de Computadores, a *World Wide Web* ("WWW" ou simplesmente Web), a *Web* Ciência ou *Web Science*, através de [Berners-Lee et al., 2006].

Os objetivos da Web Science, segundo [Fensel et al., 2011], são:

- Entender o funcionamento da Web através da análise dos componentes que lhe permitem agir como um sistema de informação descentralizado;
- Entender como cresce a Web aproveitando ou controlando esse crescimento para benefício da sociedade.

A criação da WWW por Tim Berners-Lee no final da década de oitenta previa a ligação entre documentos hipertexto e hipermídia com a finalidade de compartilhamento de informações [Berners-Lee, 1989]. No entanto, as aplicações desenvolvidas para a Web não adotaram necessariamente os padrões recomendados para tal. Cada desenvolvedor de aplicações e soluções, principalmente as comerciais, seguiu seu próprio padrão e um dos princípios da WWW, o da colaboração, não foi disseminado eficientemente. Dessa forma, o desenvolvimento da Web aconteceu permitindo o isolamento dos dados dentro das

aplicações ou repositórios remotos e dificultando, assim, a ligação semântica formal entre esses dados.

O fato é que existe um grande volume de dados de diversas áreas do conhecimento disponível na Internet através de aplicações Web. A disponibilização desses dados foi feita sem padronização e sem considerar a semântica das informações contidas nos dados. Grande parte das informações disponíveis não está estruturada e só pode ser processada por seres humanos.

Neste contexto, a Web Semântica foi idealizada como uma forma de agregar a estrutura para o conteúdo significativo das páginas disponíveis na Web, criando um ambiente onde agentes de software, ao percorrerem página por página, pudessem facilmente realizar tarefas para os usuários [Berners-Lee et al., 2001]. Em outras palavras, a Web Semântica adiciona significado às informações disponíveis na Internet de modo que o conjunto homem/máquina possa processar a semântica dos dados. Porém, para que isso ocorra, é necessária a disponibilização estruturada desses dados através de padrões definidos para a Web.

A comunidade internacional que desenvolve os padrões da Web, o W3C[1] (*World Wide Web Consortium*), define a Web Semântica como uma Web de Dados. De forma mais específica, como sendo um conjunto de tecnologias ativamente desenvolvidas para oferecer uma estrutura comum dos dados, permitindo o compartilhamento e a reutilização dos mesmos além dos domínios das aplicações, corporações e comunidades.

A estruturação dos dados pode ser realizada utilizando-se ontologias, documentos que definem formalmente as relações entre termos através de uma taxonomia e um conjunto de regras de inferência [Berners-Lee et al., 2001]. Ainda, segundo [Noy & McGuinness, 2001], servem para representar, organizar e compartilhar conhecimento sobre um determinado domínio, facilitando a gestão do conhecimento. As ontologias definem, portanto, a semântica que se quer agregar para os dados a serem disponibilizados.

O processo de construção de ontologias não é trivial e depende da colaboração entre pesquisadores que estudam ontologias e pesquisadores especialistas em cada domínio para o qual se desenvolve uma ontologia. Por

[1] http://www.w3.org

exemplo, ninguém melhor que um ornitólogo para definir a taxonomia para o domínio de aves. No entanto, chegar a um consenso sobre a representação de um determinado domínio é uma tarefa complexa e que muitas vezes não é cumprida nem por pesquisadores de uma mesma área de atuação. Um ponto fundamental no processo de estruturação é que as ontologias consolidadas sobre determinado domínio sejam reutilizadas tanto quanto possível.

A informação deve ser estruturada, ainda, de modo a ser disponibilizada em um formato que seja comum para a comunidade que irá potencialmente utilizá-la. O modelo padrão para troca de dados na Web definido pelo W3C é a Estrutura para Descrição de Recursos (*Resource Description Framework* - RDF[2]), desenvolvida para facilitar a integração de dados mesmo que pertençam a esquemas diferentes.

A integração de diferentes fontes de dados e sua publicação na Web de forma estruturada, colaborativa e intensiva podem ser alcançadas com a utilização do modelo de dados RDF, incluindo o uso de elos (*links*) RDF, tornando os dados ligados semanticamente.

O termo Dados Ligados, ou *Linked Data*, que segundo [Fensel et al., 2011] pode ser considerado um sub-tópico e uma extensão da Web Semântica, refere-se a princípios, regras e práticas adotadas para a publicação e interligação de dados estruturados na Web com a finalidade de facilitar a exploração desses dados [Berners-Lee, 2007].

Na prática, a implementação das ideias do *Linked Data* e da Web de Dados é realizada pelo projeto *Linking Open Data* (LOD)[3]. O objetivo do projeto é estender a Web através da publicação de vários conjuntos de dados abertos [Fensel et al., 2011].

A ligação de dados abertos na Web Semântica visa justamente possibilitar a integração de dados vislumbrada pela *e-Science*, objetivando viabilizar projetos colaborativos entre várias instituições e várias disciplinas.

Essa integração semântica de bases de dados científicos é um dos focos de interesse deste trabalho, na medida em que contribui para que as atividades de pesquisa em biodiversidade sejam aceleradas, aprimoradas e impactadas positivamente em médio ou longo prazo através do uso de dados integrados na

[2] http://www.w3.org/RDF
[3] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Web de Dados. Esta visão é uma adesão aos desafios e tendências indicados por [Fox & Hendler, 2009].

1.2. Cenário da pesquisa no INPA

O Instituto Nacional de Pesquisas da Amazônia - INPA, órgão federal do Ministério da Ciência, Tecnologia e Inovação com mais de cinco décadas de existência, realiza o estudo científico do meio físico e das condições de vida da região amazônica e possui a seguinte missão: "Gerar e disseminar conhecimentos e tecnologias e capacitar recursos humanos para o desenvolvimento da Amazônia" conforme [INPA, 2011].

O INPA possui vários programas de pesquisa com bases de dados científicos sobre a biodiversidade amazônica dentre os quais destacam-se: o Programa de Coleções e Acervos Científicos - PCAC[4]; o Programa de Grande Escala da Biosfera- Atmosfera na Amazônia - LBA[5]; o Programa de Pesquisa em Biodiversidade PPBio[6]; a Rede Temática de Pesquisa em Modelagem Ambiental da Amazônia - GEOMA[7]; e a Rede CTPetro Amazônia[8], uma rede temática cooperativa entre instituições de ensino superior e de ciência e tecnologia na Amazônia. Ainda, o INPA apoia em sua estrutura os quatro Institutos Nacionais de Ciência e Tecnologia - INCTs: o ADAPTA - INCT de Adaptações da Biota Aquática da Amazônia[9]; o CENBAM - INCT de Estudos Integrados da Biodiversidade Amazônica[10]; o MADEIRAS - INCT de Madeiras da Amazônia[11]; e o SERVAMB - INCT dos Serviços Ambientais da Amazônia[12].

Os dados científicos sobre coleções biológicas e outros experimentos científicos no mundo, hoje disponíveis em bases de dados, podem ser acessados via aplicações Web ou aplicações específicas, ditas aplicações orientadas a problema, apresentando impedâncias de interoperabilidade de sistemas. Além disso, as coleções de dados estão isoladas em seus domínios e a integração dessas

[4] http://www.inpa.gov.br/colecoes/colecoes2.php
[5] http://lba.inpa.gov.br/lba/
[6] http://ppbio.inpa.gov.br/
[7] http://www.geoma.lncc.br/
[8] http://projetos.inpa.gov.br/ctpetro/index.php
[9] http://adapta.inpa.gov.br/
[10] http://ppbio.inpa.gov.br/cenbam/inicio
[11] http://inctmadeiras.inpa.gov.br/
[12] http://inct-servamb.inpa.gov.br/

bases de dados entre si e com outras, ainda não é realizada nem localmente e nem via Web, o que permitiria uma colaboração mais eficiente com outros programas de pesquisa.

As iniciativas dos programas de pesquisa são muito válidas e geram conhecimento sobre a Amazônia. Porém, são investimentos e esforços similares, independentes uns dos outros e sem controle refinado dos seus objetivos devido à falta de uma política institucional de dados. E a crítica mais contundente é que os produtos finais desses programas são, na maioria das vezes, bases de dados científicas que ainda são pouco disseminadas e potencialmente pouco exploradas devido ao cenário de tecnologias adotadas.

Um fator de mitigação dos problemas citados é o mapeamento desses dados através de esquemas RDF e sua disponibilização na Web Semântica. Os esquemas RDF estruturam os dados e permitem que sejam estabelecidas ligações entre itens de diversas fontes de dados de coleções biológicas no mundo. A divulgação dos dados na LOD potencializa a colaboração das pesquisas em diferentes domínios, estimulando a reutilização de ontologias e a ligação de dados abertos. Ainda, é importante que as iniciativas atentem para as grandes e repentinas mudanças tecnológicas para que as mesmas possam ser realizadas de forma rápida e de fácil adequações, ao contrário de desenvolvimento de um novo produto com funcionalidades similares.

1.3. Motivação do trabalho

Dois dos grandes desafios da pesquisa em computação no Brasil propostos pela Sociedade Brasileira de Computação - SBC, isto é, (1) Gestão da Informação em grandes volumes de dados multimídia distribuídos e (2) Modelagem computacional de sistemas complexos artificiais, naturais e socioculturais e da interação homem-natureza, contextualizam a natureza das pesquisas, dos dados e metadados dos centros de pesquisa da Amazônia.

Certamente, um relevante momento para os grupos de pesquisa é a discussão de estruturas de conhecimento que não descartem ou pretendam substituir a especificidade das diferentes áreas do conhecimento, mas possam desenvolver metodologias inter e até mesmo transdisciplinares [Albuquerque, 2011].

Neste contexto, as investigações sobre a maneira como as diferentes bases de dados do INPA podem ser integradas e combinadas através das tecnologias da Web Semântica, bem como com outras bases espalhadas pelo mundo, convergem com os eixos temáticos definidos pelo Plano Diretor do INPA [INPA, 2011]. O interesse é proporcionar a colaboração entre grupos de pesquisas em um mesmo domínio e com outros de domínios diferentes.

Outro interesse, este relativo à gestão do conhecimento científico, é identificar metodologias de pesquisas cujos detalhes dos experimentos residem na mente de cada pesquisador (quem detém o conhecimento), implicitamente nas atividades por eles desenvolvidas (maneira pela qual o conhecimento é adquirido), que possam ser mapeadas (codificadas para disseminação) e processadas também pelo conjunto homem e máquina.

Sendo assim, a motivação do trabalho é colaborar com o processo de apoio à pesquisa em biodiversidade, através da utilização da Web Semântica, de modo a contribuir para a gestão do conhecimento científico.

Em um aspecto mais geral, a abordagem transdisciplinar do trabalho pode ajudar a responder mais rapidamente no futuro questões sociais como, por exemplo, "qual o impacto ocasionado pela vazante do Rio Amazonas com relação à ocorrência de doenças tropicais como a dengue e a malária na cidade de Manaus", supondo que seja realizada uma integração entre as bases de dados sobre as áreas de foco das doenças com as bases de dados sobre regimes fluviais.

1.4. Objetivos

O objetivo geral deste trabalho é aplicar tecnologias da Web Semântica nos processos de pesquisas e gestão de coleções de dados científicos sobre a Amazônia.

O trabalho visa contribuir com os pesquisadores do INPA nos processos de:

1. **Extração da semântica dos dados digitalizados existentes nas coleções biológicas do instituto**: As bases de dados das coleções biológicas, bem como qualquer base tradicional (dados e metadados), possuem informações que só podem ser efetivamente utilizadas mediante consulta aos pesquisadores que desenvolveram e gerenciam as bases. Tais informações, se compreendido o processo

de pesquisa, juntamente com metodologias e padronizações adotadas, podem ser agregadas às bases de dados convencionais, conferindo-lhes uma semântica que não está incorporada no modelo relacional tradicional, transformando-as no que se denominam bases de dados semânticas, permitindo a inferência de novas informações a partir daquelas existentes.

2. **Divulgação da informação na Web de Dados:** Tão importante quanto identificar e extrair as informações pertinentes ao processo de pesquisa é a divulgação dos dados semanticamente estruturados para que possam ser processados por máquinas e usuários. Neste trabalho, alguns conjuntos de dados do INPA são mapeados para bases de dados semânticas. Os dados científicos institucionais asseguram a alta qualidade de proveniência.
3. **Colaboração na pesquisa, tanto internamente quanto com outros centros de pesquisa:** Após o mapeamento para bases de dados semânticas, é fundamental, para viabilizar a colaboração entre centros de pesquisa, que os dados sejam disponibilizados na LOD. É esse compartilhamento que tornará possível a integração ou combinação entre bases de dados de mesmo domínio em uma federação ou até mesmo entre domínios diferentes que possam se inter-relacionar.

Para a composição do trabalho, destacam-se alguns objetivos específicos centrados na estruturação, manipulação e divulgação de dados científicos, bem como no desenvolvimento de aplicações para a Web Semântica, observando-se as iniciativas e tecnologias desenvolvidas como oportunidade de soluções de problemas, propiciando o fortalecimento de iniciativas regionais:

- No que se refere à estruturação dos dados, objetiva-se a construção de bases de dados semânticas considerando-se o esquema de dados CLOSi [Campos dos Santos, 2003] e a ontologia de domínio OntoBio [Albuquerque, 2011] desenvolvidos no escopo da biodiversidade amazônica, e priorizando-se ferramentas que possibilitem o acompanhamento do processo de construção, bem como a publicação dessas bases considerando-se as práticas adotadas pelo projeto LOD.

- Quanto à manipulação dos dados, objetiva-se a implementação de consultas definidas junto aos pesquisadores do INPA utilizando-se o Rexplorator, ferramenta desenvolvida na PUC-Rio para a exploração semântica de esquemas [Araújo, 2008; Azevedo, 2010].
- Outro objetivo específico é a avaliação do impacto do uso de tecnologias da Web Semântica no processo de pesquisa dos usuários. Nesta avaliação serão identificadas as vantagens e desvantagens da utilização dessas tecnologias considerando a opinião do principal interessado no processo, o pesquisador.

1.5. Organização da dissertação

A seguir, são descritos os conteúdos dos demais capítulos desta dissertação:

- **Capítulo 2 - Fundamentos**: apresentam-se discussões sobre a evolução das tecnologias utilizadas da Web estática à Web Semântica, sobre a evolução da representação de dados biológicos, sobre a evolução dos sistemas de monitoramento global da biodiversidade e, por fim, é traçado um paralelo entre a evolução Web e a evolução das coleções de dados biológicos.
- **Capítulo 3 - Metodologia**: apresentam-se os materiais e métodos utilizados para a criação das bases semânticas, para a exploração e manipulação semântica e para a construção de aplicações Web de múltiplos propósitos.
- **Capítulo 4 - Exemplos de Aplicação**: apresentam-se exemplos de aplicação da metodologia do trabalho para a Coleção de Aves da Amazônia produzida no INPA. Neste capítulo, além de apresentado o cenário da Coleção de Aves do INPA, são descritos: o processo de construção de uma base semântica utilizando-se os dados reais dessa coleção que já estão digitalizados; a manipulação dessa base através da construção de consultas definidas a partir de casos de uso elaborados junto aos pesquisadores; e, a construção de aplicação Web para o domínio da Coleção. Na última secção do capítulo, é discutida a

impressão inicial dos pesquisadores sobre aplicação de tecnologias da Web Semântica no seu domínio de pesquisa.

- **Capítulo 5 - Considerações Finais**: são apresentados os resultados alcançados e a contribuição do trabalho, bem como são sugeridos os trabalhos futuros.

2
Fundamentos

2.1.
Tecnologias: da Web Estática à Web Semântica

A *World Wide Web*, proposta por Tim Berners-Lee em março de 1989[13], foi inicialmente utilizada para o compartilhamento e interligação de documentos na rede mundial por uma comunidade restrita de usuários, em sua maioria composta por pesquisadores.

Em pouco mais de duas décadas, com a utilização em massa da WWW pela sociedade para os mais variados fins, acompanhou-se a evolução do sistema inicialmente proposto com limitações para uma das tecnologias mais complexas do mundo moderno.

A WWW é alicerçada desde o seu início por três pilares [Fensel et al., 2011]:

1. HTTP (*HyperText Transfer Protocol*) - um protocolo padrão para a recuperação de hipertextos e outros documentos por sistemas de informação hipermídia distribuídos e colaborativos;
2. URI (*Uniform Resource Identifier*) - cadeias de caracteres (*strings*) utilizadas para identificar exclusivamente recursos através da Internet. Existem dois tipos de URI: o URN (*Uniform Resource Name*) que define a identidade de um recurso; e o URL (*Uniform Resource Locator*) que define a localização de um recurso. A sintaxe para definir um URI é:
    - *scheme : [// authority][/ path][? query][# fragid],* onde, *scheme* é usado para distinguir os diferentes tipos de URI, *authority* é normalmente identificado como um servidor, *path* é identificado como um diretório ou arquivo no servidor, *query* adiciona parâmetros extras e *fragid* identifica um recurso secundário.

[13] http://www.w3.org/History/1989/proposal.html

3. HTML (*HiperText Markup Language*) - um formato utilizado para a criação e publicação de documentos hipertextos interligados.

O HTTP se baseia no mecanismo de nomeação do URI e fornece uma maneira de publicar e recuperar recursos descritos utilizando HTML.

Hipertextos podem ser vistos como uma forma de construir documentos que se referem a outros recursos da Web, não necessariamente outros documentos, mas tudo o que possa ser identificado por um URI. A Web em si, é o maior exemplo de implementação de hipertexto.

No início da Web, os documentos hipertexto eram estáticos, pois referenciavam recursos estáticos, ou seja, recursos que não possuem alteração de conteúdo ou estrutura. A WWW poderia ser denominada, então, de Web Estática. Os usuários assumiram o papel de consumidores do conteúdo disponível em documentos descrevendo, por exemplo, catálogos ou coleções de dados de referência cruzada.

Posteriormente, os documentos hipertexto passaram a referenciar também recursos dinâmicos, como em um portal de notícias, permitindo o desenvolvimento de sistemas de ligação mais complexos e dinâmicos. Neste momento, a WWW poderia ser denominada de Web Dinâmica. Os usuários mantiveram o papel de consumidores, porém, de conteúdo dinâmico disponibilizado em uma estrutura de ligação mais complexa.

Em 2007, Nova Spivack[14] propôs uma linha do tempo para a evolução da Web, apresentada na Figura 1.

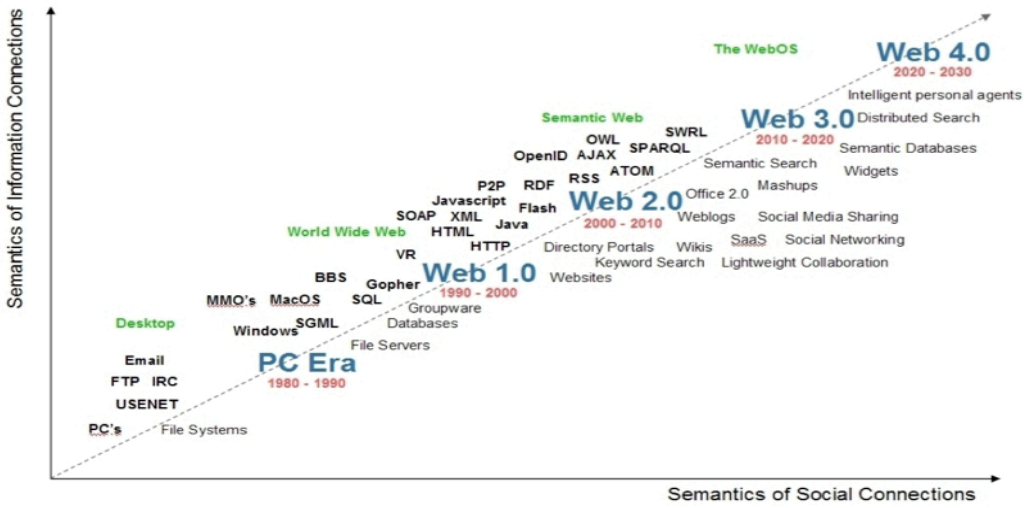


Figura 1 - Linha do tempo para a evolução da Web

[14] http://www.novaspivack.com/technology/how-the-webos-evolves

Spivack sugere que a evolução da Web ocorra com base na relação entre a riqueza semântica das conexões sociais e a riqueza semântica das conexões de informação. Em uma relação diretamente proporcional, a Web evolui quanto mais semanticamente ricas forem as conexões sociais e as conexões de informação.

A Web Estática e a Web Dinâmica estariam representadas na Figura 1 como a Web 1.0, na qual a maioria dos esforços foram concentrados para o desenvolvimento de tecnologias voltadas para a publicação da informação.

A evolução para Web 2.0, a Web atual, significou, mais que um avanço tecnológico, uma mudança no foco de desenvolvimento da tecnologia. A mídia deixou de ser desenvolvida para o indivíduo e passou a ser desenvolvida para a comunidade; e os usuários, antes consumidores de conteúdo e de serviços da WWW, passaram a ser também produtores.

Ainda segundo [Fensel et al., 2011], apesar de todo o avanço tecnológico em prol da Web, ainda existem algumas limitações que precisam ser superadas visando à continuidade da sua evolução:

- localização de informação relevante - as consultas para recuperação de informação são baseadas em palavras-chave e são feitas em linguagem natural escrita. Com isso, é comum o retorno de resultados que contém termos homônimos mas com significado irrelevante para o usuário, bem como a ausência de resultados com termos sinônimos e que deveriam ser recuperados;
- extração de informação relevante das páginas Web - a extração da informação é feita através de *wrappers*[15], mas não é escalável, pois depende do formato e *layout* dos sítios Web;
- combinação e o reuso da informação disponível na Web - no processo de busca pela informação, a mesma consulta é refeita várias vezes.

Para lidar com essas limitações, foi proposta a Web Semântica, uma solução que permite às máquinas "entenderem" e potencialmente satisfazerem as consultas do usuário, através do processamento do significado da informação.

[15] De acordo com a definição de [Fensel et al., 2011], *wrappers* são unidades de *software* genéricas que separam a implementação dos componentes do Ambiente de Execução do Serviço Web do mecanismo de comunicação. Eles fornecem métodos para os componentes, permitindo a comunicação com outros componentes. *Wrappers* são gerados automaticamente e anexados a cada componente durante a instanciação.

2.1.1.
Web Semântica

> A Web Semântica não é uma nova Web, mas uma extensão da atual, na qual a informação é disponibilizada com um significado bem definido, provendo melhor colaboração entre computadores e pessoas.
>
> [Berners-Lee et al., 2001]
>
> A visão da Web Semântica é estender os princípios da Web de documentos para dados. Dados deveriam ser acessados utilizando a arquitetura geral da Web utilizando, por exemplo, URIs; dados deveriam ser relacionados uns com os outros assim como documentos (ou partes de documentos) já o são. Isso também significa a criação de um *framework* comum que permita que dados sejam compartilhados e reutilizados através dos domínios das aplicações, corporações e comunidades, para serem processados automaticamente por ferramentas tão bem quanto manualmente, incluindo a criação de possíveis novos relacionamentos entre as peças de dados.
>
> [Herman, 2001]

Assim como a Web de Documentos, a Web de Dados é composta por documentos Web. Porém, enquanto na Web convencional as ligações são escritas em HTML e relacionam documentos hipertexto, na Web de Dados as ligações relacionam "coisas" arbitrárias descritas em RDF.

A Figura 2 ilustra como são estabelecidas as ligações na Web de Documentos e na Web de Dados, de acordo com [Shadbolt, 2010].

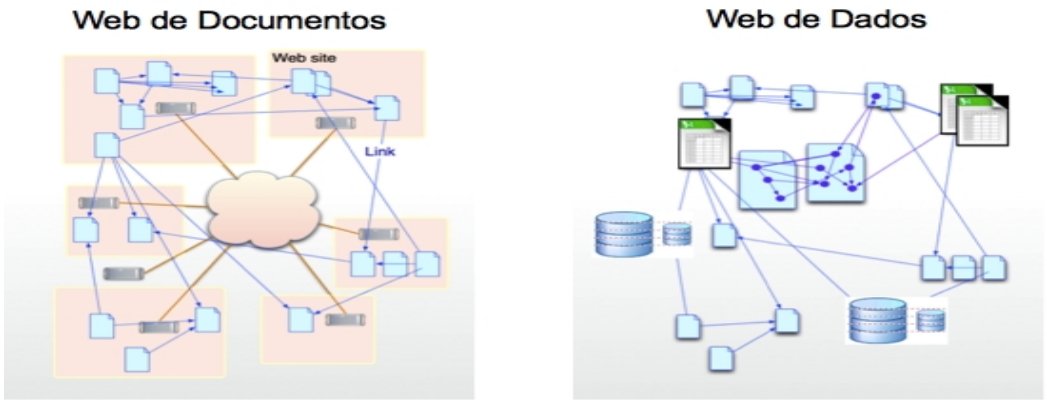


Figura 2 - As ligações na Web de Documentos e na Web de Dados

Segundo Berners-Lee[16], a implementação da Web Semântica depende diretamente da divulgação estruturada de dados na Web e da ligação desses dados uns com os outros de modo que tanto as pessoas quanto as máquinas possam explorar a Web de Dados.

[16] http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee definiu quatro regras ou princípios básicos a serem seguidos para a publicação de dados na Web Semântica e que são os fundamentos *Linked Data*:

1. Utilizar URIs como nomes para as "coisas", não somente documentos Web ou conteúdo digital, mas também objetos do mundo real e conceitos abstratos;
2. Utilizar HTTP URIs de modo que as pessoas possam procurar por aqueles nomes e receber alguma informação como resultado da consulta, ou seja, permitir que estes URIs sejam dereferenciáveis;
3. Fornecer informações úteis utilizando os padrões da Web Semântica (RDF, por exemplo) quando alguém acessar um URI;
4. Incluir *links* para outros URIs de forma que eles possam localizar informações adicionais.

A adoção das práticas de *Linked Data* permite a criação de *links* entre diferentes fontes de dados e, assim, a conexão destas fontes em um único espaço global de dados. O uso de padrões Web e um modelo comum de dados torna possível a implementação de aplicações genéricas que operam por todo o espaço global.

As subseções a seguir apresentam as melhores práticas *Linked Data* adotadas para a construção da Web Semântica de acordo com [Heath & Bizer, 2011].

A utilização de HTTP URIs como nomes para as coisas

A publicação de dados na Web pressupõe inicialmente a identificação dos itens de interesse de um domínio, denominados como "recursos" na arquitetura Web [Jacobs & Walsh, 2004]. As propriedades e relacionamentos destes recursos é que serão descritos em dados. E os HTTP URIs permitem a criação de bons nomes por dois motivos:

1. Eles fornecem uma maneira simples de criar nomes globais únicos de forma descentralizada;
2. Além de servirem como nome, eles são um meio de acessar informações sobre o recurso identificado.

A criação de URIs dereferenciáveis

Todo HTTP URI deve ser dereferenciável, ou seja, clientes HTTP procuram por um URI utilizando o protocolo HTTP e recuperam uma descrição do recurso identificado pela URI. Esta prática permite que tanto URIs utilizados para identificar documentos HTML quanto URIs utilizados no contexto *Linked Data*, identifiquem objetos do mundo real e conceitos abstratos.

As descrições dos recursos são incorporados como documentos Web. Descrições a serem lidas por humanos devem ser representadas em HTML e as descrições a serem lidas por máquinas devem ser apresentadas em RDF.

Se os URIs identificarem objetos do mundo real, é fundamental não confundir os objetos em si com os documentos Web que os descrevem. A prática a ser adotada é utilizar diferentes URIs para identificar o objeto do mundo real e o documento que o descreve, evitando assim a ambiguidade. Esta prática permite separar sentenças feitas sobre um objeto e sobre um documento que descreva aquele objeto.

O Modelo de Dados RDF e *Linked Data*

De acordo com [Manola & Miller, 2004], a descrição de um recurso em RDF é representada em forma de triplas <*sujeito, predicado, objeto*>. Uma tripla espelha a estrutura básica de uma afirmação como <*Dendrocygna autumnalis*, *temNomeVulgar*, *asa-branca*>. O sujeito de uma tripla é o URI de identificação do recurso descrito. O objeto pode ser tanto um valor literal, quanto o URI de um outro recurso que está relacionado ao sujeito. O predicado, no meio da tripla, indica qual tipo de relação existe entre o sujeito e o objeto. O predicado também é identificado por um URI.

Ocorrem dois tipos de triplas RDF:

1. Triplas Literais que possuem um literal RDF, como um *string*, um número ou uma data, e são usados para descrever as propriedades dos recursos;
2. *Links* RDF que descrevem o relacionamento entre dois recursos e são compostos por três referências URI. Os URIs nas posições de sujeito e objeto no link identificam seus respectivos recursos. O URI na posição de predicado define o tipo de relacionamento entre os recursos. Pode-

se distinguir *links* RDF entre internos e externos. *Links* RDF internos conectam recursos dentro de uma única fonte *Linked Data*, com a presença dos URIs do sujeito e do objeto no mesmo *namespace*, ou seja, no mesmo espaço identificador. *Links* RDF externos conectam recursos que são servidos por diferentes fontes *Linked Data*, com a presença dos URIs do sujeito e do objeto em diferentes *namespaces*.

Um conjunto de triplas RDF pode ser visto como um grafo[17] RDF composto por nós e arcos representando os recursos, suas propriedades e seus valores. Os URIs ocorrendo como sujeito e objeto são os nós no grafo e cada tripla é um arco direcionado que conecta o sujeito e o objeto.

A seguir são listados os principais benefícios do uso do modelo de dados RDF no contexto das práticas *Linked Data*:

1. O modelo RDF é inerentemente projetado para ser usado em escala global e permite a qualquer um se referir a qualquer coisa através do uso de HTTP URIs como identificadores globais únicos para itens de dados, bem como para termos de vocabulário;
2. Clientes podem encontrar qualquer URI em um grafo RDF através da Web e obter informação adicional;
3. O modelo de dados permite que *links* sejam estabelecidos entre dados de diferentes fontes;
4. Informações de diferentes fontes podem ser facilmente combinadas através da mescla de dois conjuntos de triplas em um único grafo;
5. O RDF permite a representação de informações de diferentes esquemas em um único grafo; isto significa que podem ser misturados termos de diferentes vocabulários para representar dados;
6. Combinado a linguagens de esquema como o *Resource Description Framework Schema* (RDFS) [Brickley & Guha, 2004] e a *Ontology Web Language* (OWL) [McGuinness & Van Harmelen, 2004], o modelo de dados permite o uso de estrutura, tanto ou tão pouco quanto o desejado, o que significa que podem ser representados dados fortemente estruturados, bem como dados semi-estruturados.

[17] http://pt.wikipedia.org/wiki/Grafo

Apesar dos benefícios descritos, algumas funcionalidades do modelo de dados RDF devem ser evitadas no contexto das práticas *Linked Data*:

1. A reificação RDF, uma vez que sentenças reificadas são de difícil implementação em SPARQL, linguagem de consulta a bases de dados em RDF [Prud'hommeaux & Seaborn, 2008].
2. As coleções e os contêineres RDF [Antoniou & Van Harmelen, 2008] também são problemáticos se os dados tiverem que ser consultados com SPARQL e, nos casos em que a ordem relativa de itens em um conjunto não é significativa, o uso de múltiplas triplas com o mesmo predicado é recomendado;
3. O uso de nós anônimos ("*blank nodes*"), pois não é possível definir *links* RDF externos para eles e a mescla de dados de diferentes origens fica muito difícil quando nós anônimos são usados.

Formatos de Serialização RDF

Para se publicar um grafo RDF na Web é necessário primeiro serializá-lo utilizando uma sintaxe RDF, ou seja, transcrever suas triplas para um arquivo utilizando-se uma sintaxe particular.

A sintaxe RDF/XML [Beckett, 2004; Manola & Miller,2004] é padronizada pelo W3C e amplamente usada para divulgar dados na Web. Porém, por ser considerada de difícil leitura e escrita pelos seres humanos, não é recomendada para gestão e curadoria de dados que envolvem intervenção humana. O Quadro 1 apresenta uma serialização RDF/XML para duas triplas.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/">

   <rdf:Description rdf:about="http://biglynx.co.uk/people/dave-smith">
      <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
      <foaf:name>Dave Smith</foaf:name>
   </rdf:Description>

</rdf:RDF>
```

Quadro 1 - Serialização RDF/XML [Heath & Bizer, 2011]

A primeira indica que existe uma coisa identificada pelo URI *http://biglynx.co.uk/people/dave-smith* do tipo *Person*; e a segunda indica que esta coisa tem o nome *Dave Smith*.

Outro formato de serialização padronizado pelo W3C, o RDFa (*Resource Description Framework in Attributes*) [Adida & Birbeck, 2008] incorpora triplas RDF em documentos HTML entrelaçando dados RDF dentro do modelo de objeto de documento (*Document Object Model - DOM*). O RDFa é utilizado por editores de dados que modificam os modelos HTML, mas não têm familiaridade com a infraestrutura de publicação. O Quadro 2 apresenta a serialização RDFa para as mesmas triplas do Quadro 1.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
"http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:foaf="http://xmlns.com/foaf/0.1/">

   <head>
      <meta http-equiv="Content-Type" content="application/xhtml+xml;
         charset=UTF-8"/>
      <title>Profile Page for Dave Smith
   </head>

   <body>
      <div about="http://biglynx.co.uk/people#dave-smith"
         typeof="foaf:Person">
         <span property="foaf:name">Dave Smith
      </div>
   </body>

</html>
```

Quadro 2 - Serialização RDFa [Heath & Bizer, 2011]

Além dos formatos de serialização padronizados, existem os formatos de texto simples, utilizados com o mesmo propósito, como o *Turtle* [Beckett & Berners-Lee, 2008] que suporta prefixos de *namespaces* e outros atalhos. É um formato mais amigável para leitura de triplas RDF e escrita de triplas à mão. O Quadro 3 apresenta a serialização com *Turtle* para o mesmo par de triplas do Quadro 1.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://biglynx.co.uk/people/dave-smith>
rdf:type foaf:Person ;
foaf:name "Dave Smith" .
```

Quadro 3 - Serialização Turtle [Heath & Bizer, 2011]

Outro formato não padronizado é o N-Triples[18], um subconjunto de *Turtle*. O N-Triples possui menos recursos, não suportando prefixos de *namespaces* e

[18] http://www.w3.org/TR/rdf-testcases/#ntriples

nem os atalhos, o que resulta em um formato de serialização com redundância. Esta redundância gera arquivos maiores que os de *Turtle* ou mesmo o RDF/XML, mas possibilita que os arquivos N-Triples sejam analisados (*parsed*) mesmo que não possam ser alocados em memória principal e possibilita uma melhor compressão, reduzindo o tráfego de rede na troca de arquivos. O Quadro 4 apresenta a serialização com N-Triples para o mesmo par de triplas do Quadro 1.

```
<http://biglynx.co.uk/people/dave-smith>   <http://www.w3.org/1999/02/22-
rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .

<http://biglynx.co.uk/people/dave-smith> <http://xmlns.com/foaf/0.1/name>
"Dave Smith" .
```

Quadro 4 - Serialização com N-Triples. Fonte: [Heath & Bizer, 2011]

A inclusão de *links* para outras fontes de dados

O quarto princípio das práticas *Linked Data* é incluir ligações RDF apontando para outras fontes de dados na Web. Esses *Links* externos RDF são cruciais para a Web de Dados pois funcionam como elemento essencial que liga as ilhas de dados em um espaço global de dados completamente interligados.

Tecnicamente, um *link* RDF é uma tripla RDF em que o sujeito da tripla é uma referência URI no *namespace* de um conjunto de dados, enquanto o predicado e/ou o objeto da tripla são referências URIs apontando para os *namespaces* de outros conjuntos de dados. A dereferência desses URIs produz uma descrição do recurso fornecido pelo servidor remoto. Esta descrição geralmente contém *links* RDF adicionais que apontam para outros URIs que, por sua vez, também podem ser dereferenciados, e assim por diante. Esta é a maneira como as descrições dos recursos individuais são tecidas na Web de Dados. Esta é também a forma como a Web de dados pode ser explorada usando um navegador *Linked Data* ou rastreada por robôs de mecanismos de busca.

Existem três tipos importantes de links RDF:

1. Links de Relacionamento - apontam para coisas relacionadas, mas disponíveis em diferentes fontes de dados. Por exemplo, links de relacionamento permitem que pessoas apontem para informações relacionadas ao lugar em que vivem, ou a dados bibliográficos sobre as publicações que elas escreveram.
2. Links de Identidade - apontam para "sinônimos" (*aliases*) do URI usados por diferentes fontes de dados para identificar o mesmo objeto

do mundo real ou seu conceito abstrato. Links de identidade permitem aos usuários obter mais descrições sobre uma entidade disponíveis em outras fontes de dados. Ainda, possuem uma importante função social porque permitem que visões diferentes do mundo se expressem na Web de Dados.

3. Links de Vocabulário - apontam a partir dos dados para as definições dos termos do vocabulário que são usados para representar os dados, bem como destas definições para as definições de termos relacionados e presentes em outros vocabulários. Links de vocabulário tornam os dados auto-descritivos e permitem aos aplicativos *Linked Data* compreenderem e integrarem dados através de vocabulários.

Esquema de classificação 5 estrelas

Antes da adoção das práticas *Linked Data*, os dados estruturados eram comumente disponibilizados na Web através de formulários, publicados como depósito de dados de *Comma-Separated Values* (CSV), planilhas, diversos formatos de dados específicos de domínio e passaram também a ser disponibilizados via Web *Application Programming Interfaces* (APIs).

A arquitetura criada com as práticas *Linked Data* facilitou aos consumidores descobrir, acessar e integrar dados. O processo para publicação de dados de acordo com as práticas *Linked Data* é contínuo no sentido de tornar mais fácil consumir e trabalhar com os dados

Berners-Lee[19] descreve essa continuidade através de um esquema de 5 estrelas, segundo o qual os publicadores de dados podem atribuir estrelas aos seus conjuntos de dados de acordo com o seguinte critério:

- 1 Estrela: quando os dados estão disponíveis na Web em qualquer formato, mas com uma licença aberta;
- 2 Estrelas: quando os dados estão disponíveis na Web como dados estruturados legíveis por máquina;
- 3 Estrelas: quando os dados estão disponíveis na Web como dados estruturados legíveis por máquina e em formato aberto;

[19] Berners-Lee também disponibiliza a descrição do sistema 5 estrelas em http://www.w3.org/DesignIssues/LinkedData.html

- 4 Estrelas: além de estarem disponíveis conforme as classificações anteriores, quando os dados seguem padrões abertos definidos pelo W3C (RDF e SPARQL);
- 5 Estrelas: além de estarem disponíveis conforme as classificações anteriores, quando os dados são ligados aos de outras pessoas para fornecer contexto.

Fundamentalmente, cada classificação pode ser obtida por vez, o que representa uma transição progressiva para *Linked Data*, ao invés de uma adoção completa em uma única operação.

2.2. A evolução da representação de dados biológicos

A representação de dados biológicos apresenta-se como um antigo desafio para a ciência da computação uma vez que esses dados apresentam um nível de complexidade elevado incluindo parâmetros espaço-temporais, estrutura indefinida, multidimensionalidade, vocabulário incógnito expresso por uma linguagem particular, grande volume de dados e evolução dinâmica [Albuquerque, 2011].

Nos anos 70s, foram publicadas as bases do modelo relacional o que permitiu a representação dos dados com base na modelagem conceitual e a partir de um modelo de entidades e relacionamentos, o modelo ER [Chen, 1976]. Juntamente com este modelo era necessário manter o Dicionário de Dados, um documento com a definição de todos os objetos criados no modelo.

O que de fato o Dicionário de Dados representa é uma coleção de metadados contendo definições e representações de todos os elementos de dados do modelo. Esses metadados são dados que descrevem a estrutura dos dados, suas restrições, suas aplicações, suas autorizações, e assim por diante [Elmasri & Navathe, 2010].

Em meados dos anos 80s, surgiram os modelos de dados orientados a objeto e os modelos de dados objeto-relacionais [Rumbaugh, 2006]. este momento, a representação de dados ganhou em semântica, mas ainda não era capaz de expressar com fidelidade sistemas com alto grau de complexidade, por exemplo, sistemas de biodiversidade.

O domínio da biologia possui uma grande diversidade e variabilidade de conceitos herdados da complexidade dos sistemas biológicos. Além disso, os dados biológicos são repletos de exceções devido à evolução desses sistemas e do progresso tecnológico [Macêdo, 2005].

Neste contexto e já na década de 90, as ontologias surgiram como uma alternativa para tentar suprir essa limitação de expressividade semântica dos modelos de representação de dados até então existentes. Segundo [Albuquerque, 2011], as ontologias passaram a ser empregadas em diversas áreas, como Inteligência Artificial (IA), Engenharia de Software e Web Semântica, possibilitando a criação de modelos conceituais claros, concisos e não ambíguos.

Durante a evolução da representação de dados biológicos foram sendo desenvolvidos modelos de representação de dados sobre a biodiversidade. Dentre eles, merecem destaque o *Darwin Core*, o *Access to Biological Collection Data* (ABCD), *o Plinian Core*, o *Ecological Metadata Language* (EML) e o *Clustered Object Schema for INPA's Biodiversity Data Collections* (CLOSi) que serão apresentados nas subseções a seguir.

2.2.1. *Darwin Core*

O *Darwin Core*[20] [Blum & Wieczoreck, 2005] é um padrão criado para facilitar o compartilhamento de informações sobre a diversidade biológica, fornecendo definições de referência, exemplos e comentários. É baseado principalmente nas categorias taxonômicas, as suas ocorrências na natureza, conforme documentado por meio de observações, espécimes e amostras, e informações relacionadas.

Esse padrão é baseado nas normas desenvolvidas pela *Dublin Core Metadata Iniciative*[21] (DCMI) e deve ser visto como uma extensão do *Dublin Core*, proposto inicialmente para metadados de obras impressas e objetos digitais em geral (por exemplo, título, criador, data, assunto), para informações sobre a biodiversidade. Os termos descritos neste padrão fazem parte de um conjunto

[20] http://rs.tdwg.org/dwc/index.htm
[21] http://dublincore.org/

maior de vocabulários e especificações técnicas em desenvolvimento e mantidos pelo *Biodiversity Information Standards*[22] (TDWG).

O *Darwin Core* é composto por um glossário de termos e se propõe fornecer definições semânticas estáveis objetivando a máxima reutilização em vários contextos.

2.2.2. *Access to Biological Collection Data* (ABCD)

O ABCD[23], também mantido pelo TDWG, é um esquema XML comum de especificação de dados para unidades de coleta biológicas, incluindo espécimes vivos e preservados, juntamente com observações de campo que não produzem amostras para a coleção. O esquema suporta a troca e a integração de dados de coleções biológicas.

Todas as coleções biológicas do mundo contém uma série de itens de dados incluindo elementos específicos para os espécimes (por exemplo, categoria taxonômica, sexo, etc.) e para a coleção (por exemplo, a instituição que realizou a coleta do dado). O conjunto de elementos utilizados varia de coleção para coleção. O ABCD fornece um conjunto reconciliado de nomes de elementos e suas definições para uso de cientistas e curadores.

Objetivando uma especificação de dados mais abrangente e geral, bem como a imposição mínima de elementos requeridos de modo a tornar a especificação funcional, o esquema não abrange dados taxonômicos, como sinonímia, ou relacionados à categoria taxonômica, como intervalo de distribuição e valores indicadores.

Os elementos e conceitos que são usados proveem o máximo possível de compatibilidade com o outros padrões no campo de coleções de dados biológicas, como o *Darwin Core*.

[22] http://www.tdwg.org/
[23] http://wiki.tdwg.org/twiki/bin/view/ABCD/WebHome

2.2.3.
Plinian Core

O *Plinian Core*[24] é um padrão sobre espécies biológicas projetado para ser utilizado no processo automatizado de integração e de recuperação de informação de bases de dados heterogêneas.

Esse padrão é desenvolvido pelo *Sistema Costarricense de Información sobre Biodiversidad*[25] (CRBio) e pelo *Global Biodiversity Information Facility in Spain*[26] (GBIF.ES), nós da rede *Global Biodiversity Information Facility*[27] (GBIF).

O objetivo do padrão é especificar os conceitos básicos para integrar e recuperar informação sobre espécies de organismos distribuídas em bases de dados por instituições em todo o mundo, visando apoiar tomadores de decisão, pesquisadores da biodiversidade, professores, estudantes, formadores de opinião, produtores da economia sustentável, e o público em geral.

2.2.4.
Ecological Metadata Language (EML)

A EML[28] [McCartney & Jones, 2002] é uma especificação de metadados desenvolvido pela disciplina de ecologia e para a disciplina de ecologia. É implementada como um conjunto de documentos XML que podem ser utilizados de forma modular e extensível para documentar dados ecológicos. Cada módulo EML é projetado para descrever uma parte lógica do total de metadados que devem ser incluídos em qualquer conjunto de dados ecológicos.

As primeiras versões da linguagem foram desenvolvidas no *National Center for Ecological Analysis and Synthesis* (NCEAS), na Universidade da Califórnia, nos Estados Unidos. A versão atual da EML é desenvolvida pelo Projeto EML que é composto totalmente por membros voluntários em prol do avanço do gerenciamento de informação para a ecologia.

[24] http://www.pliniancore.org/
[25] http://crbio.cr/site/index.html
[26] http://www.gbif.es/
[27] http://www.gbif.org/
[28] http://knb.ecoinformatics.org/software/eml/

O objetivo do Projeto EML é fornecer uma especificação de metadados de alta qualidade, em código aberto, para descrever dados relevantes para a disciplina de ecologia.

2.2.5. *Clustered Object Schema for INPA's Biodiversity Data Collections* (CLOSi)

Em 2003, Campos dos Santos [Campos dos Santos, 2003] apresentou o *Clustered Object Schema for INPA's Biodiversity Data Collections* (CLOSi), um esquema conceitual de banco de dados para representação das coleções biológicas do INPA que é base para uma visão integrada dos dados destas coleções.

O esquema é constituído por seis *clusters* integrados (*Colleciton Management, Taxonomy, Reference, Collecting Event Of Collection, Locality Of Biodiversity Data, Agent Of Collection*) e descrito por um conjunto de classes de objetos, complementados por classes de valores controlados de objetos inter-relacionados.

O CLOSi foi desenvolvido para facilitar e estimular o desenvolvimento dos bancos de dados das coleções biológicas do INPA, mas beneficia institutos similares e pode ser considerado como um padrão de biodados e metadados.

Este esquema serve como base para a criação de ontologias para as coleções de dados biológicos do INPA e é revisto na secção 3.2.1 do Capítulo 3.

2.3. Evolução de sistemas para gestão de dados e informações de biodiversidade em ambiente integrado

A biodiversidade pode ser representada por uma variedade de dados, isto é, registros de espécies, dados geográficos, ecológicos, socioeconômicos, etc. Segundo [Albuquerque, 2011], muitos desafios foram elencados neste contexto, dentre eles:

- a identificação e avaliação de potencial interrupção do fluxo de conhecimento essencial da biodiversidade, o que inclui os aspectos taxonômicos e de distribuição geográfica;

- o delineamento de experimentos cujo planejamento seja eficiente no que se refere ao levantamento e descrição dos organismos em grupos de extrema importância ou sob ameaça de extinção;
- a localização e o mapeamento de dados biológicos, essencialmente de coleções; e
- a concepção de novas abordagens para a utilização das informações de maneira integrada.

Tais desafios adquirem maior complexidade quando dados de diferentes comunidades que trabalham com multidomínios requerem integração. Algumas iniciativas já apresentam indícios de êxito, tendo como um dos resultados o grande volume de informações geradas que exigem soluções não triviais para gestão, análise e síntese [Albuquerque & Campos dos Santos, 2005].

Sistemas de Informação de Biodiversidade (SIBs) manipulam dois tipos fundamentais de informação: 1) registros de catálogos e acervos de museus (exemplo de material depositado no Programa de Coleções do INPA) e 2) registros que documentam coletas e observações feitas em campo (mantidas em diversos meios de armazenamentos). Em ambos os casos, as informações descrevem espécies: identificação e classificação taxonômica, período das coletas, local, a metodologia e os agentes envolvidos no processo de coleta. Devido ao processo de gestão e manutenção de coleções, muitas informações que poderiam ser compartilhadas são mantidas repetidas em sistemas específicos, promovendo multiplicidade (redundâncias desconectas), dificuldade na integridade e integração de dados. Um exemplo de registro de coleta pode ser observado na Figura 3.

Figura 3 - Exemplo de documentos da coleta da classe Arachnida UFAM/INPA [Bonaldo et al., 2009]

Na Figura 3, os dados não são estruturados e o documento apresenta uma estrutura simples. Não se pode afirmar que *I. limbata* faz referência ao gênero e espécie do organismo coletado e *Anyphaenidae* à família. Para pesquisadores ou coletores tal duvida não existe uma vez que conhecem o domínio e natureza dos dados, no entanto, o mesmo não é verdadeiro para pesquisadores de comunidades e domínios científicos diferentes.

O propósito dos SIBs é auxiliar pesquisadores a aprimorarem ou complementarem seu conhecimento e entendimento sobre os seres vivos [Torres et al., 2006]. A exploração desses dados se concentra em quando e onde foram observados, por quem e qual processo utilizado, e informações geográficas, caracterizando os ecossistemas onde os espécimes foram observados ou coletados, além da distribuição espacial das ocorrências nos biomas.

A demanda por SIBs para avaliar as questões ambientais, por exemplo, espécies ameaçadas de extinção, desmatamento, capacidade hidrológica para uso na malha energética nacional, e bioprospecção para fármacos e cosméticos, está em constante crescimento. Uma grande questão, ainda sem resposta, é: tem-se a informação desejada? Informação pode existir, mas o problema reside em como obtê-la. Os veículos de divulgação de resultados científicos, em geral, não apresentam todas as informações necessárias. Desta forma, as coleções biológicas desempenham papel imprescindível no atendimento de demandas e na tentativa de

responder complexos questionamentos, uma vez que coleções representam esforços delineados de experimentos e anos de investigação sobre a fauna, flora, macro e microbiota.

A computação tem sido um recurso fundamental no gerenciamento de informações de biodiversidade. Sua utilização está associada a algumas demandas [Albuquerque & Campos dos Santos, 2005]: um modelo de informações precisas, gerenciamento de dados formais e padrões de metadados, bem como métodos para integrar e revitalizar dados legados (preparação para análises refinadas).

2.3.1. SIBs e suas aplicações para ambiente integrado

Um grande número de projetos de desenvolvimento de SIBs tem como objetivo disponibilizar funcionalidades para gerenciar e publicar dados disponíveis na Web. Um exemplo consolidado é o SpeciesLink[29]. Este sistema Web integra a informação de coleções biológicas e observações documentadas, depositadas em museus, herbários e coleções microbiológicas, integrando-as via protocolo específico para em seguida publicar na Web.

Outro exemplo de sucesso é o Specify[30], que adota uma plataforma computacional que utiliza serviços Web como suporte para o gerenciamento das coleções de dados, incluindo descrição geográfica da coleta, dados dos coletores e funções para gerência operacional da coleção.

Outros SIBs são os programas desenvolvidos para gerenciar dados de coletas de campo. Um exemplo é o projeto Biota, que propôs um sistema de banco de dados para gerenciar inventários de biodiversidade para o projeto ALAS (*Artropodos de La Selva*) [Colwell, 1996; Biota, 2010]. O sistema SinBiota gerencia registros de observações de campo realizadas por grupos de pesquisa financiados pela FAPESP, Fundação de Amparo à Pesquisa do Estado de São Paulo[31].

Projetos de abrangência global, como o GBIF, *Integrated Taxonomic Information System*[32] (ITIS), *Species 2000*[33], o TDWG, entre outros, buscam

[29] http://splink.cria.org.br/
[30] http://www.specifysoftware.org/Specify/
[31] http://www.cria.org.br/
[32] http://www.itis.gov/
[33] http://www.sp2000.org/

estabelecer aplicações e padrões para a integração e a interoperabilidade de dados das coleções biológicas e disseminação na Web. O GBIF é uma organização internacional cujo objetivo é disponibilizar informação sobre biodiversidade em rede distribuída de bancos de dados interoperáveis, respeitando a propriedade intelectual dos fornecedores de dados.

Uma característica comum das aplicações de biodiversidade é a sua concentração no nível taxonômico de espécies. Isso ocorre porque as espécies são a base de um sistema de agrupamento hierárquico conhecido como árvore taxonômica, usado pelos cientistas para classificar formas de vida [Morris et al., 2007]. Assim, outro conjunto considerável de sistemas de biodiversidade lida com o gerenciamento de informações taxonômicas e a distribuição geográfica das espécies. Como exemplo, *The Tree of Life* [Maddison & Schulz, 2007], *Catalogue of Life*[34], OBIS-SEAMAP [Halpin, 2006], e TaiBIF [Shao, 2007]. O projeto *The Tree of Life* é um esforço internacional para prover informação sobre a diversidade de organismos na terra, suas características e evolução histórica. O projeto *Catalogue of Life* visa fornecer um catálogo mundial de taxonomia das espécies vivas unificando essa informação em um sistema de banco de dados que seja mundialmente acessível. O projeto OBIS-SEAMAP é um banco de dados com referência espacial para coleções de espécies marinhas que podem ser visualizadas usando aplicações que apresentam mapas. O TaiBIF integra a informação de biodiversidade de Taiwan, compreendendo lista de espécies, imagens, informações geográficas, informação ambiental, informação disponível na literatura, informação fornecida por especialistas e uma lista de instituições e organizações relevantes. O comum entre todos esses projetos é a utilização da Web como mecanismo de disseminação da informação.

Outra abordagem encontrada na literatura são ferramentas que permitem a identificação de espécies baseadas no conceito de guias de campo, um livro elaborado para ajudar na identificação de espécies. Por exemplo, *Electronic Field Guide* (EFG), uma ferramenta que permite aos cientistas redigir e gerar suas próprias guias de campos e sofisticadas chaves de identificação taxonômica, que podem ser publicadas e compartilhadas na Internet [Morris et al., 2007].

[34] http://www.catalogueoflife.org/

2.4.
Paralelo entre evolução Web e evolução das coleções biológicas

A Web evolui visando o compartilhamento de informação em geral e as coleções biológicas evoluem visando a geração de conhecimento científico sobre as ciências da vida.

Desde o seu início, a Web evoluiu com a participação dos seus usuários. Se em um primeiro momento os usuários eram meros consumidores de informação, assumiram logo em seguida o papel de divulgadores individuais de conteúdo e já começam a atuar como colaboradores na geração compartilhada de conhecimento. O ambiente compartilhado que caracteriza a Web favoreceu o crescimento rápido e a evolução anunciada para um espaço global de dados.

As coleções biológicas, por sua vez, evoluíram isoladamente e dependendo de usuários especialistas.

No caso dos institutos da Amazônia, apesar dos conjuntos de dados científicos serem abundantes, são incompatíveis e dispersos. Isto conduziu ao desenvolvimento de sistemas espontâneos nos quais as aplicações são dirigidas por problemas específicos sem levar em conta a necessidade de integração de dados, escalabilidade da arquitetura do sistema e disseminação da informação [Campos dos Santos, 2003].

O momento atual, com a eminência da Web Semântica, favorece as coleções biológicas pois possibilita a disponibilização dos dados científicos, até então isolados nas soluções orientadas a problema, no espaço global. Esta disponibilização permitirá o desenvolvimento colaborativo de soluções por usuários especialistas em uma nova realidade de fontes de dados heterogêneas, acessadas via Web, com escala e conteúdo consistentes, compatíveis e integrados.

A Figura 4 ilustra esse novo cenário em potencial para as coleções biológicas dos institutos da Amazônia, através da evolução da *Figura 2.7 - Sources of biodiversity data and information* presente em [Campos dos Santos, 2003].

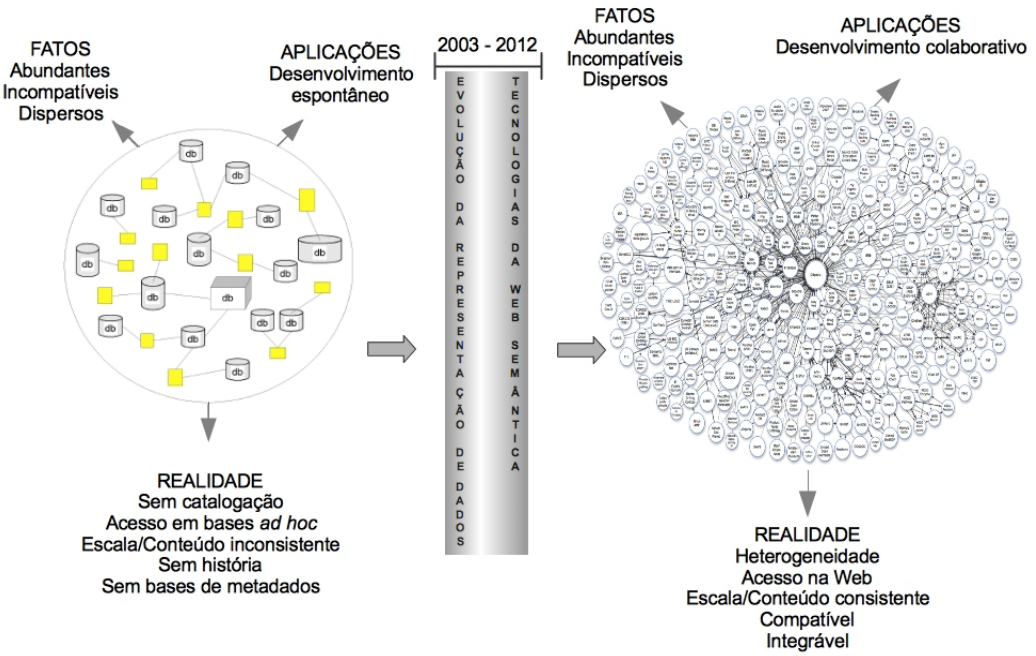


Figura 4 - Evolução do cenário das fontes de dados e informações sobre a biodiversidade amazônica

3
Metodologia

3.1.
Introdução

> A Metodologia é o estudo dos métodos. Ou as etapas a seguir num determinado processo.
>
> Tem como objetivo captar e analisar as características dos vários métodos indispensáveis, avaliar suas capacidades, potencialidades, limitações ou distorções e criticar os pressupostos ou as implicações de sua utilização.
>
> […]. A metodologia é também considerada uma forma de conduzir a pesquisa ou um conjunto de regras para o ensino de ciência.
>
> A Metodologia é a explicação minuciosa, detalhada, rigorosa e exata de toda ação desenvolvida no método (caminho) do trabalho de pesquisa. É a explicação do tipo de pesquisa, dos instrumentos utilizados (questionário, entrevista etc), do tempo previsto, da equipe de pesquisadores e da divisão do trabalho, das formas de tabulação e tratamento dos dados, enfim, de tudo aquilo que se utiliza no trabalho de pesquisa.
>
> […]
>
> Metodologia refere-se a mais do que um simples conjunto de métodos, refere-se aos fundamentos e pressupostos filosóficos que fundamentam um estudo particular.
>
> [Wikipédia[35]]

Este Capítulo apresenta a metodologia utilizada para o desenvolvimento deste trabalho. Por se tratar da aplicação de tecnologias utilizadas no contexto da Web Semântica nas coleções biológicas do INPA e compreender várias fases, também são citados os trabalhos relacionados a cada uma delas.

3.2.
Criação das bases de dados semânticas

Bases de dados semânticas permitem às máquinas processarem o significado da informação nelas contida.

Estas bases viabilizam o acesso à informação de forma mais inteligente, porque permitem que agentes automatizados entendam e "decodifiquem" informações muito mais rápido que o ser humano. Basicamente, eles dividem a

[35] http://pt.wikipedia.org/wiki/Metodologia

informação em sua forma mais simples para que seja rápida e facilmente compreendida pelo usuário.

Os dados são organizados em triplas da forma <sujeito, propriedade, valor> sendo interpretados de forma significativa, sem intervenção humana, em modelos binários de objetos. Bases de dados criadas em torno deste conceito têm maior aplicabilidade e são mais facilmente integradas a outras bases de dados [Hull & King, 1987].

O propósito das bases de dados semânticas é permitir aos computadores compreender e descobrir informações sem a ajuda do homem. Isso significa que o computador seria capaz de ter um processo de pensamento complexo e resolveria problemas por conta própria. Ao invés de ser programado para cada tarefa em particular, o computador seria capaz de programar-se com base em experiências passadas, encontrar informações por conta própria, combiná-las com outras informações, conforme necessário, e agir sobre a informação recebida de forma adequada.

3.2.1. Esquema de dados e ontologia de domínio: pontos de partida

A construção de bases de dados semânticas no contexto de dados sobre biodiversidade pode fazer uso de um esquema conceitual de dados do domínio como base para a modelagem semântica e de regras de nomenclatura zoológica para compor parte do vocabulário.

O objeto de estudo deste trabalho envolve dados de coleções biológicas e, intrinsecamente, suas rotinas de coletas de campo, curagem, análise e visualização para a gestão do conhecimento científico.

O projeto de uma base de dados para gerenciamento de coleções biológicas demanda a compreensão de cada uma dessas atividades. Requer também conhecimento dos dados e suas características. Para tanto, usuários especialistas devem estar envolvidos nos processos de identificação de requisitos de dados e do sistema, especialmente durante a análise de requisitos de dados.

No contexto do INPA, esta fase de análise foi conduzida através de entrevistas com os próprios pesquisadores, da coleta de documentos técnicos internos e da avaliação das descrições. Para agregar valor semântico aos dados, é importante que cada participante no processo seja especialista em algum grupo

taxonômico ou em certo aspecto biológico de algum grupo taxonômico. As entrevistas podem apresentar um formato aberto e ser compostas pelas mesmas questões gerais dirigidas aos pesquisadores. Pesquisas na Web e bibliográficas devem complementar as fontes de estudos utilizadas como subsídios para a geração das bases de dados semânticas.

Tipicamente, entende-se conceitualmente que as instituições possuem coleções biológicas compostas por objetos coletados em determinada localidade. Estes objetos podem ser classificados segundo uma taxonomia específica referenciada em trabalhos científicos.

> Os dados coletados durante uma missão de campo, para registro de espécies, são de dois tipos: os gerais, que constituem informações que são normalmente coletadas em todos os estudos (por exemplo, dia, hora, descrição da localidade), e os específicos, que correspondem ao interesse científico de um estudo (exemplo, altitude de uma localidade ou a fase da lua podem ser de interesse de um estudo, mas não de outro). Entrevistar cientistas que trabalharam em diferentes estudos e em diferentes áreas ajuda na diferenciação entre informações comuns a todos e aquelas utilizadas apenas por poucos cientistas.
>
> [Campos dos Santos, 2003]

Como mencionado na secção 2.2.5, Campos dos Santos apresentou, em 2003, um esquema conceitual para representação das coleções biológicas do INPA chamado CLOSi.

CLOSi é o resultado de estudos realizados em conceituadas instituições científicas que desenvolvem pesquisas no âmbito da Amazônia a saber: o INPA; a Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA); o Instituto de Pesquisas Científicas e Tecnológicas do Estado do Amapá (IEPA); o Museu Paraense Emílio Goeldi (MPEG); e, o Laboratório de Silvicultura (Silvolab) na Guiana Francesa.

Em 2011, Albuquerque apresentou uma ontologia de biodiversidade, OntoBio, baseada no CLOSi e em requisitos levantados no INPA atendendo ao protocolo de coleta de dados de biodiversidade [Albuquerque, 2011].

A utilização de esquemas de dados, ontologias e dos dados disponíveis na nuvem LOD, bem como suporte ao processo de desenvolvimento de bases de dados semânticas, torna-se viável atualmente, pois se trata do uso de uma conceitualização já concebida do domínio da aplicação, reuso de vocabulários e

interligação de dados publicados nesta nuvem para a construção de uma nova conceitualização com nível de detalhamento diferenciado.

A seguir, são apresentados os principais recursos do esquema de dados CLOSi e da OntoBio que, pelo fato de terem sido criados para o contexto do cenário biológico da Amazônia, constituem material relevante a ser considerado no desenvolvimento deste trabalho.

CLOSi para Informações sobre Biodiversidade

CLOSi compreende 6 *clusters* (*Collection_Management*, *Taxonomy*, *Reference*, *Collecting_Event_Of_Collection,* *Locality_Of_Biodiversity_Data*, *Agent_Of_Collection*), onde cada um é descrito por um conjunto de classes de objetos, complementados por classes de valores controlados de objetos inter-relacionados. A Figura 5 apresenta a estrutura dos grupos de conceitos inter-relacionados de coleções biológicas (*clusters*).

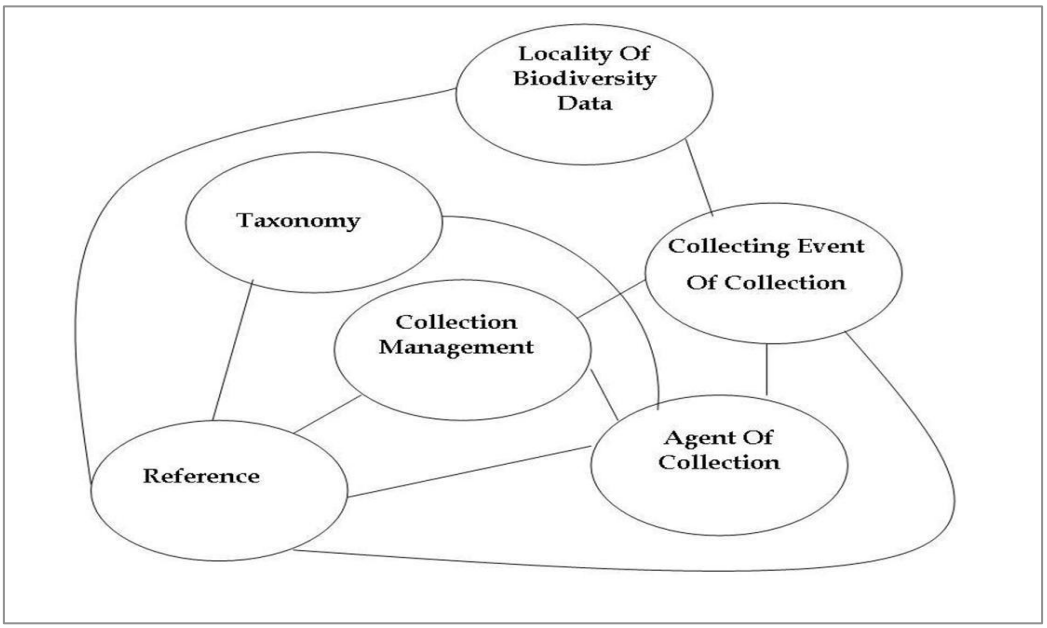


Figura 5 - Clusters e estrutura dos relacionamentos do esquema CLOSi [Campos dos Santos, 2003]

Os conceitos dos *clusters* de objetos inter-relacionados do esquema foram definidos com base nos conceitos desenvolvidos pela *Association of Systematic Collections*[36] (ASC) e pelo *Object-Protocol Model* (OPM) [Chen & Markowitz, 1995]. Estes conceitos foram estendidos, visando suportar os requisitos funcionais identificados no cenário do INPA, através de entrevistas, materiais solicitados,

[36] http://siarchives.si.edu/collections/siris_arc_217612

fluxo de dados e avaliação de descrições. Os pesquisadores participaram como usuários e os curadores das coleções como gerentes de informação e provedores de dados. A estrutura final do CLOSi foi desenvolvida a partir de uma pesquisa aprofundada das necessidades dos usuários de dados de coleções biológicas.

O esquema abrange a maioria dos aspectos gerais de dados biológicos, uma vez que seu projeto conceitual foi originado de múltiplas fontes. O fato do esquema possuir definição sintática própria e classes de valores controlados, contribui para a utilização do CLOSi como suporte e base inicial para a modelagem de bases de dados semânticas para o domínio de coleções biológicas.

Ontologia de Biodiversidade - OntoBio

A Ontologia de Biodiversidade é baseada na engenharia de ontologias e foi desenvolvida com base na pesquisa de Guizzardi [Guizzardi, 2005], conforme apresentada em [Albuquerque, 2011; Albuquerque et al., 2012].

Tal como na engenharia de software, Albuquerque apresenta uma visão inicial de análise da OntoBio como modelo conceitual em OntoUML[37], utilizando o método SABIO [Falbo, 2004]; e, posteriormente, a implementação da ontologia utilizando as duas versões da *Ontology Web Language*[38] (OWL/OWL2). Para a validação da OntoBio, foram respondidas Questões de Competência utilizando a *Semantic Web Rule Language*[39] (SWRL). A OntoBio encontra-se disponível em <http://lis.inpa.gov.br/biodiversityontology>.

O principal objetivo da OntoBio é prover uma conceitualização clara e precisa dos aspectos considerados em coletas de dados de biodiversidade independentes de uma aplicação específica. Os requisitos base da ontologia refletem este propósito e os usos esperados para ela, isto é, a competência da ontologia.

A OntoBio (Figura 6) está dividida em cinco sub-ontologias, complementares umas às outras, conectadas por relações entre seus conceitos e por axiomas formais :

1. Sub-Ontologia Coleta;
2. Sub-Ontologia Entidade Material;

[37] http://en.wikipedia.org/wiki/OntoUML
[38] http://www.w3.org/2004/OWL/
[39] http://www.w3.org/Submission/SWRL/

2.1. Sub-Ontologia Entidade Biótica;

2.2. Sub-Ontologia Entidade Abiótica;

3. Sub-Ontologia Localização Espacial;
4. Sub-Ontologia Ecossistema;
5. Sub-Ontologia Ambiente.

Os axiomas respondem às questões de competência citadas anteriormente, a fim de permitir:

- uma rica expressividade semântica que não pode ser alcançada apenas com o uso do modelo gráfico,
- as inferências (pela codificação da ontologia),
- uma avaliação da fidedignidade do apresentado com o propósito da ontologia,
- validação da ontologia, e
- identificação de inconsistências.

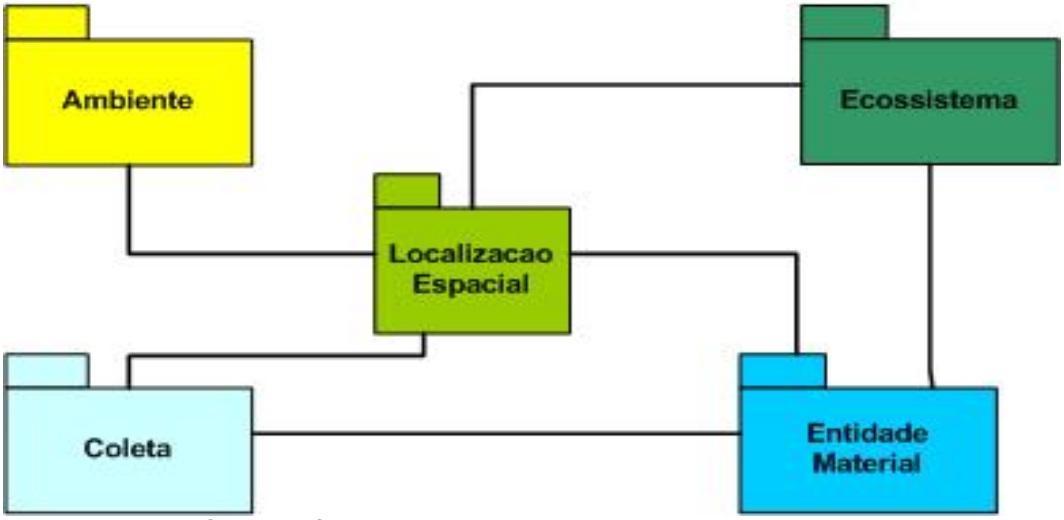


Figura 6 - Visão Geral da OntoBio

3.2.2. Etapas do processo

Considerando o CLOSi e a OntoBio como ponto de partida para a criação das bases de dados semânticas, o processo de construção destas bases e sua preparação para consumo via Web Semântica seguem as etapas da metodologia proposta pelo Grupo de Trabalho LinkedDataBR - GT-LinkedDataBR da Rede Nacional de Ensino e Pesquisa - RNP em [GT-LinkedDataBR, 2011], a saber:

1. Coleta: nesta etapa, entrevistas são realizadas com os curadores das coleções de dados biológicos do INPA com o propósito de coletar os conjuntos de dados digitalizados que possam ser utilizados no trabalho, bem como entender a dinâmica da construção e utilização dessas coleções.
2. Modelagem: nesta etapa, é realizada uma modelagem conceitual dos conjuntos de dados coletados na primeira etapa, considerando-se o CLOSi. Esta modelagem só será necessária se os conjuntos de dados ainda não possuírem um modelo de dados. Também deve ser criado o de banco de dados com base no esquema gerado pela modelagem conceitual. Este banco de dados será alimentado com os dados selecionados e extraídos na Etapa 3.
3. Seleção e extração: nesta etapa, os dados de interesse para a composição das bases de dados semânticas são selecionados e extraídos dos conjuntos de dados coletados junto aos pesquisadores.
4. Curagem: nesta etapa, os dados são curados de modo a normalizar o seu formato e tratar seu conteúdo garantindo a qualidade desses dados e evitando inconsistências ou ambiguidades semânticas na base gerada.
5. Mapeamento: nesta etapa, é utilizada uma ontologia para capturar parte do domínio semântico no qual os dados adquirem significado. Deve-se considerar a OntoBio para essa agregação de semântica aos dados das bases iniciais e transformação dos mesmos em bases de triplas, ou deve ser construída uma ontologia baseada na mesma.
6. Ligação: nesta etapa, as bases semânticas criadas são ligadas a outras existentes disponibilizadas na WS.
7. Armazenamento: nesta etapa, as triplas geradas e ligadas são armazenadas em um repositório de bases de dados semânticas para posterior consumo por aplicações da Web Semântica, como o Rexplorator e o Synth.

A construção das bases semânticas pode ser realizada através de processo de *Extraction, Transformation, Loading*[40] (ETL) que permite a integração de dados de diferentes fontes. O processo ETL é especificado a partir de *Jobs* e *Transformations*[41]. *Transformations* são tarefas orientadas a dados cuja finalidade

[40] http://pt.wikipedia.org/wiki/Extract,_transform,_load
[41] http://wiki.pentaho.com/display/EAI/Carte+User+Documentation

são a extração, a transformação e a carga dos dados. Enquanto os *Jobs* são coleções de *Transformations* orientados a tarefas.

3.2.3. Ferramentas selecionadas

Existem atualmente várias ferramentas para utilização no processo de construção de bases semânticas. O Grupo de Trabalho LinkedDataBR - GT-LinkedDataBR sugere a adoção de uma ferramenta para o gerenciamento do fluxo de criação de repositórios de dados abertos (bases de dados semânticas) utilizando os padrões de *Linked Data*. A ferramenta em questão é o *Pentaho Data Integration* - Kettle[42] que provê uma interface gráfica intuitiva para a construção de processos ETL, o *Spoon*.

Ainda em [GT-LinkedDataBR, 2011], foi realizado o levantamento e uma avaliação do estado da arte sobre ferramentas utilizadas para: a conversão de formatos, XML e dados em planilhas, para RDF; o armazenamento das bases de triplas RDF; e a interligação dos dados convertidos com outros publicados na nuvem LOD.

Em [Sahoo et al., 2009], também é apresentado um estudo sobre ferramentas e linguagens consideradas como o estado da arte para o mapeamento de bases de dados relacionais para bases em RDF.

A seleção das ferramentas a serem utilizadas na construção das bases semânticas neste trabalho é realizada de acordo com etapas da metodologia e os critérios adotados para a seleção de ferramentas são, principalmente, a licença de uso sob a qual a ferramenta está disponibilizada, seu grau de maturidade das funcionalidades e a documentação disponível sobre a mesma.

Inicialmente, a modelagem e criação de um banco de dados para a coleção biológica do INPA é realizada utilizando-se o *MySQL Workbench*[43], ferramenta de código aberto, com alto grau de maturidade e vasta documentação disponível sobre a mesma.

A seleção e extração dos dados, bem como o pré-processamento dos mesmos para o mapeamento para as bases semânticas são implementados com o

[42] http://kettle.pentaho.com
[43] http://dev.mysql.com/doc/index-gui.html

Kettle[44]. A ferramenta de código aberto é utilizada para gerenciamento de fluxos de processos, possui grau de maturidade alto e excelente documentação.

Grande parte das informações sobre as coleções biológicas do INPA são digitalizadas em planilhas. A curagem dos dados presentes nessas planilhas é realizada com o *Google Refine*[45], ferramenta utilizada na limpeza e tratamento de dados em tabelas. Ainda, a ferramenta possui uma extensão RDF que permite o mapeamento dos dados da tabela para o formato RDF via interface gráfica. A ferramenta é gratuita, com grau de maturidade desconhecido e, apesar de não possuir documentação disponível, possui vídeos tutoriais disponíveis na Web e que auxiliam na sua utilização.

Na fase de mapeamento, é construída uma ontologia, considerando-se a OntoBio, utilizando-se o *Protege*[46], editor de ontologias desenvolvido na Escola de Medicina da Universidade de Stanford, nos Estados Unidos. A ferramenta é de código aberto, apresenta alto grau de maturidade e excelente documentação disponível na Web. A ferramentas selecionadas para a fase de mapeamento são o próprio Kettle, por facilitar a continuidade dos processos de seleção e extração dos dados no mesmo ambiente de desenvolvimento, e o *D2R Server*, ferramenta desenvolvida pela *Freie Universität-Berlim*, para a publicação de bancos de dados relacionais para RDF, quando for o caso. Esta última é de código aberto, com médio grau de maturidade e suporta bem as conversões realizadas no trabalho.

As ferramentas selecionadas para a ligação dos dados são o Silk[47] e o LIMES[48] que permitem a ligação entre as triplas de dados RDF geradas no processo ETL e os conjuntos de triplas da LOD, tais como DBPedia e da *British Broadcasting Corporation*[49] (BBC). Esta ligação acontece quando há a equivalência entre entidades das triplas RDF geradas e as presentes na LOD, como por exemplo, dois recursos referindo-se à mesma espécie, à mesma localização, etc. Silk é a ferramenta selecionada para ligar os dados, seguindo o recomendado em [GT-LinkedDataBR, 2011], visto que apresenta boa documentação, estabilidade, independência de domínio e configuração avançada via arquivos XML, além de ser de código aberto. O LIMES, por sua vez, é

[44] http://www.pentaho.com/
[45] http://code.google.com/p/google-refine/
[46] http://protege.stanford.edu/
[47] http://www4.wiwiss.fuberlin.de/bizer/silk/
[48] http://aksw.org/Projects/LIMES.html
[49] http://www.bbc.co.uk/

utilizado como opção alternativo ao Silk, em casos de inatividade no servidor desta, sendo gratuito, apresentando documentação razoável, exemplo de aplicação e até mesmo uma versão online para utilização na Web[50].

E, por fim, para o armazenamento das triplas, é selecionado o *framework* de código aberto Sesame, desenvolvido pela *Aduna/OpenRDF*[51] para armazenamento, inferência e consulta RDF. A ferramenta possui alto grau de maturidade e desempenho estável e rápido, além de ser o repositório utilizado pelo Rexplorator.

3.2.4. Trabalhos relacionados

Existem vários trabalhos que tratam da criação de bases de dados nos mais diversos domínios para a publicação na nuvem LOD. O W3C mantém uma lista de estudos de caso e casos de uso da Web Semântica e, dentre eles, vale ressaltar o Estudo de Caso da BBC [Raimond et al., 2010].

Esse estudo de caso trata da ligação de Programas da BBC, cuja finalidade é informar, educar e entreter, com fontes de dados de diversos domínios do conhecimento. Mais especificamente, o buscador *BBC Wildlife* disponibiliza um identificador Web para cada espécie (e outras categorias biológicas), habitat e adaptação de interesse da BBC. A *BBC Nature* agrega dados de diferentes fontes, inclusive da *Wikipedia*, do *World Wild Fund* (WWF), do *International Union Conservation of Nature* (IUCN) sobre espécies ameaçadas de extinção, entre outras. O buscador *BBC Wildlife* reorganiza esses dados e coloca-os no contexto da BBC, ligando-os para clipes de programas extraídos do acervo BBC.

Outra iniciativa importante que vale ser mencionada é o Serviço de Dados Semânticos[52] da *European Environment Agency*[53] (EEA). O serviço compreende um motor de busca orientado a objetos através do qual se pode pesquisar o conteúdo de dados da *European Environment Information and Observation Network*[54] (Eionet), rede de informações sobre o meio ambiente europeu que disponibiliza parte de seus dados em RDF.

[50] http://limes.aksw.org/
[51] http://www.openrdf.org/
[52] http://semantic.eea.europa.eu/
[53] http://www.eea.europa.eu/
[54] http://www.eionet.europa.eu/

3.3. Construção das consultas

Uma vez criadas as bases de dados semânticas, primeiro objetivo específico deste trabalho, é necessária a construção de consultas a essas bases juntamente com os pesquisadores, especialistas em seus domínios de atuação. As consultas são desenvolvidas no Rexplorator, ferramenta para manipulação de esquemas RDF desenvolvida na PUC-Rio. Nesta subsecção, também são apresentados as demais ferramentas e trabalhos relacionados à consulta e manipulação de dados semânticos.

3.3.1. O Rexplorator

O Rexplorator foi implementado na dissertação de mestrado de Marcelo Cohen de Azevedo [Azevedo, 2010] com base no Explorator desenvolvido na dissertação de Samur Araújo [Araújo, 2008].

A ferramenta foi desenvolvida sob o paradigma *Model-View-Controller* (MVC) [Reenskaug, 1979] que define a separação de uma aplicação em três camadas fundamentais:

1. modelo - camada responsável pela lógica de negócio;
2. controle - camada que responde as requisições, trata os dados e acessa a camada de modelo;
3. visualização - camada responsável pela exibição dos resultados para o usuário da aplicação.

O Rexplorator utiliza o framework *ActiveRDF* [Oren et al., 2007] que provê uma camada de acesso aos dados representados em RDF e que suporta diversos tipos de armazenagens, como SPARQL *endpoints* ou bancos de dados RDF, como o Sesame.

A ferramenta foi desenvolvida na linguagem dinâmica *Ruby* [Flanagan & Matsumoto, 2008], o que permite ao *framework* criar classes e métodos que refletem as classes e atributos existentes no modelo RDFS das bases acessadas, abstraindo de certa forma o modelo RDF em que os dados estão representados. Com isso o desenvolvimento de aplicações pode ser feito de forma mais ágil.

Além disso, a ferramenta utiliza requisições e atualizações da interface via Ajax[55], visando uma melhor interação com o usuário.

O modelo

As aplicações desenvolvidas através da ferramenta são armazenadas em RDF em uma base interna do Rexplorator. Para isso, Azevedo definiu uma ontologia na qual todos os dados de domínio pudessem ser representados. Dessa forma, eles podem ser persistidos para que depois sejam reutilizados, alterados e compartilhados.

Todas as classes de modelo do Rexplorator herdam de RDFS::Resource, classe do *ActiveRDF*, *framework* que cuida da persistência das instâncias na base RDF. Os Atributos herdados de RDFS:Resource são *uri* e *class_uri*. O primeiro define a URI do recurso, uma espécie de identificador único na base RDF interna do Rexplorator. O segundo define a URI da classe a que recurso pertence.

O modelo do Rexplorator é apresentado no diagrama apresentado na Figura 7.

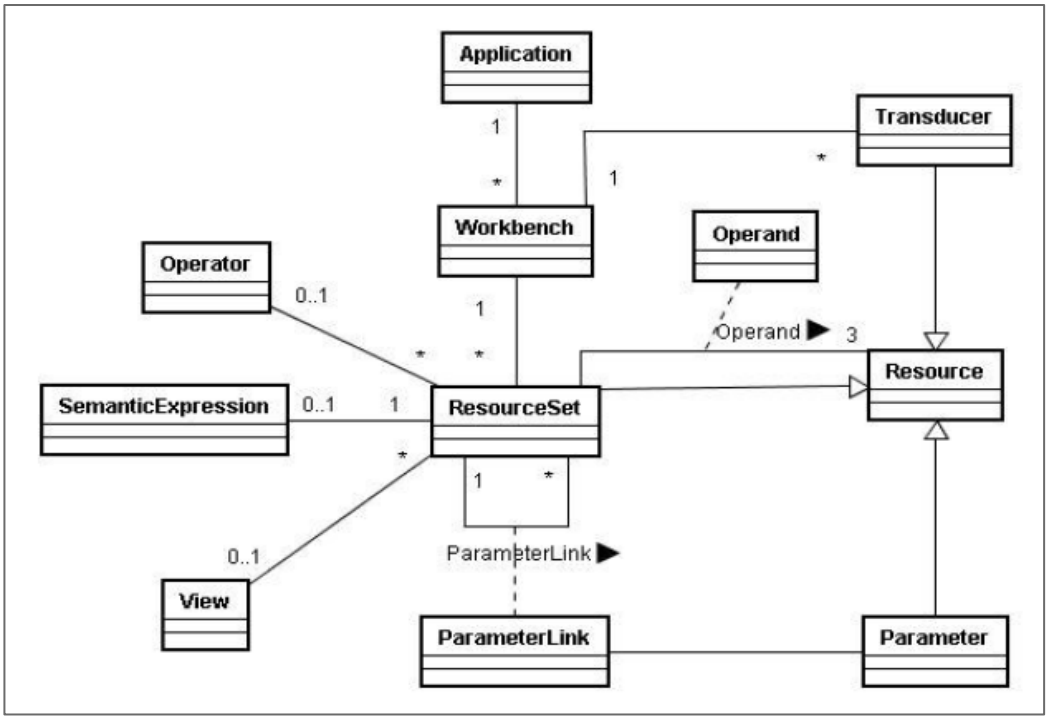


Figura 7 - Modelo do Rexplorator [Azevedo, 2010]

A seguir são descritas resumidamente cada uma das classes do modelo apresentado por Azevedo em sua dissertação de mestrado.

[55] Asynchronous Javascript And Xml - http://ajaxpatterns.org/

A classe *Application*

A classe ***Application*** representa uma aplicação construída no Rexplorator. Usuários podem persistir aplicações e carregar aplicações previamente persistidas a qualquer momento. No momento da persistência, é possível escolher um novo nome para a aplicação. Além de um nome, aplicações possuem um dono, um conjunto de operações habilitadas e um conjunto de *workbenches*.

A classe *Workbench*

O recurso da classe ***Workbench*** representa um conjunto de consultas que pretende responder a uma pergunta específica, por exemplo, quais espécies de aves de um determinado Estado estão presentes na Coleção de Aves da Amazônia.

Cada aplicação possui sempre no mínimo um *workbench* e assim como sempre existe uma aplicação ativa para cada usuário utilizando o sistema, sempre existe também um *workbench* ativo. Recursos da classe ***Workbench*** possuem nome, descrição, um indicador que define se o *workbench* pode ser copiado por outros usuários, uma coleção de transdutores e uma coleção de conjuntos de triplas.

A classe *ResourceSet*

A classe ***ResourceSet*** representa um conjunto de triplas RDF. As triplas de um *resourceset*, instância de ***ResourceSet***, são um subconjunto das triplas existentes nas bases RDF habilitadas no Rexplorator. Instâncias da classe ***ResourceSet*** são criadas pelo usuário e as suas triplas são definidas de duas formas: ou através de uma expressão que define uma consulta SPARQL às bases habilitadas; ou através de uma operação em cima de outros recursos. Um *resourceset* criado diretamente com uma expressão que define uma consulta SPARQL não pode ser modificado depois de sua criação. Já um *resourceset* criado a partir de uma operação em cima de outros recursos pode ser editado e modificado. Um *resourceset* desse tipo possui um operador e operandos. O operador é definido no momento de criação do *resourceset* e não pode ser modificado. Já os operandos são coleções de recursos e podem ser alterados.

A classe *Operation*

A classe de domínio ***Operation*** representa uma operação sobre um ou mais conjuntos de triplas. O resultado da operação é sempre o conjunto de triplas de um *resourceset*. As operações podem ser criadas dinamicamente por usuários e compartilhadas. O código que define o comportamento da operação é um atributo da classe ***Operation***, e deve ser um trecho de código na linguagem *ruby*. Uma instância da classe ***Operation*** também possui um nome e um atributo com o símbolo da operação, que deve ser um caractere que represente a operação na ferramenta.

A classe *Operand*

A classe ***Operand*** existe para adicionar atributos no relacionamento entre um conjunto de triplas e seus operandos. Quando um conjunto de triplas tem como um de seus operandos outro conjunto de triplas, ele necessita de um atributo extra para indicar qual projeção deve se considerar como seu parâmetro. A projeção é uma das posições das triplas: sujeito, predicado ou objeto. A classe ***Operand*** possui um atributo que adiciona esse metadado ao relacionamento. Por exemplo, considerando um conjunto de triplas A que é definido por uma operação SPO e possua como único parâmetro outro conjunto de triplas B para a posição do sujeito de A. Caso a projeção de B seja sujeito, serão utilizados os recursos que estão na posição do sujeito nas triplas de B; caso seja predicado, serão utilizados os recursos que estão na posição de predicado nas triplas de B e caso seja objeto, serão os recursos que estão na posição de objeto nas triplas de B. Portanto, é necessário guardar qual a projeção de B será utilizada na operação.

A classe *Parameter*

A classe ***Parameter*** facilita a substituição de valores que compõem os operandos de um resourceset, possibilitando a generalização de consultas e a reutilização. Um recurso da classe ***Parameter*** é criado quando o usuário seleciona um recurso de um dos operandos de um resourceset para ser parametrizado. Um parâmetro possui um nome, que é gerado pelo sistema no momento da sua criação, e um tipo de entrada de dados. Ambos podem ser editados pelo usuário.

A classe *ParameterLink*

Recursos da classe ***ParameterLink*** pertencem a um resourceset e são utilizados pelo compartilhamento fechado da aplicação. Quando um resourceset possui um ParameterLink, o recurso selecionado pelo usuário final da aplicação é enviado como valor de um parâmetro para um outro resourceset. Para isso, recursos da classe ***ParameterLink*** possuem uma instância da classe ***Parameter***, a instância da classe ***ResourceSet*** a qual o parâmetro pertence e uma projeção (sujeito, predicado ou objeto). Quando um recurso R de um conjunto de triplas A é selecionado por um usuário na aplicação fechada e existe alguma instância de ParameterLink de A que possua a projeção a que R pertence, o valor do recurso R será enviado para ser avaliado na posição do parâmetro indicado pela instância do ParameterLink.

A classe *Transducer*

A classe de domínio ***Transducer*** permite a entrada de dados pelo usuário que não se encontram nos repositórios RDF habilitados e que podem ser utilizados pelas operações na formação de conjuntos de triplas. Os Tansdutores do Rexplorator possuem um nome e um valor, ambos editáveis via texto livre.

A classe *View*

A classe ***View*** permite a customização da interface, utilizada em aplicações fechadas, e possui como atributos um nome e o código fonte em HTML customizado. Usuários da ferramenta podem criar e alterar instâncias da classe ***View*** e dessa forma alterar a aparência e o comportamento da camada de apresentação.

3.3.2. Trabalhos relacionados e outras ferramentas

Existem outras formas de explorar bases de dados semânticas. Nesta subsecção, são apresentadas outras ferramentas utilizadas com este propósito, bem como os trabalhos relacionados ao Rexplorator.

Uma maneira alternativa ao Rexplorator para se realizar consultas a bases de dados semânticos é através de SPARQL, linguagem de consulta declarativa,

recomendada pelo W3C para extrair informações de grafos RDF com base em correspondência de padrões de tripla. No entanto, para se utilizar diretamente a SPARQL é necessário conhecer a sintaxe da linguagem, além das ontologias utilizadas na geração das bases semânticas. É possível utilizar consultas SPARQL para a compreensão das ontologias nas quais os dados estão representados. Sendo SPARQL uma linguagem de consulta de baixo nível, consultas pouco complexas são relativamente simples de serem montadas. Porém, construir consultas mais complexas que relacionem e filtrem diversos dados se mostra uma tarefa trabalhosa, difícil e pouco eficiente.

Outra forma de acessar o conteúdo semântico contido em bases RDF é através do projeto e implementação de uma aplicação, para uma base ou ontologia específica, através de *frameworks* que minimizam ao máximo o retrabalho do desenvolvedor nas partes comuns de acesso aos dados . Dentre esses *frameworks*, destacam-se o o *Jena*[56] e o *ActiveRDF* [Oren et al., 2007]. Porém, a Web Semântica necessita de ferramentas genéricas que permitam a construção de aplicações específicas de forma fácil e extensível.

Existem também ferramentas que permitem somente a consulta a bases RDF de forma visual para facilitar a exploração dos dados, como o *Paged Graph Visualization* (PGV) [Deligiannidis et al., 2007] e o *gFacet* [Heim et al., 2008], e outras que permitem transformações em cima dos dados das bases , como o DERI Web Data Pipes[57] e o Explorator [Araújo, 2008].

[56] http://jena.sourceforge.net/
[57] http://pipes.deri.org

4
Exemplo de Aplicação

4.1.
Introdução

Na década de 90, o INPA estabeleceu o Programa de Coleções e Acervos Científicos - PCAC, coordenado por um Gerente de Coleções e organizado em estruturas funcionais denominadas curadorias, para cada uma das coleções botânicas (Herbário, Carpoteca e Xiloteca), zoológicas (Anfíbios e Répteis, Aves, Invertebrados, Mamíferos) e Microbiológicas (de Interesse Agrossilvicultural e de Interesse Médico), que têm a responsabilidade de manter, gerenciar e desenvolver as coleções.

Essas coleções representam parte significativa do conhecimento sobre a biodiversidade amazônica gerado no instituto. Elas mantém representantes dessa diversidade biológica em conservação ex-situ, isto é, fora do lugar de origem, vivos ou fixados, cujo o público-alvo são pesquisadores e estudantes de pós-graduação da sociedade acadêmica nacional e internacional.

Ainda, segundo o Regimento Interno do PCAC atualizado [INPA, 2006], os bancos de dados eletrônicos das coleções representam extensões lógicas das coleções científicas biológicas do INPA e da sua documentação física, constituindo parte integrante dessas coleções.

Os dados do Herbário, por exemplo, são registrados em uma versão institucional do *Botanical Research And Herbarium Management System* - BRAHMS[58] e os dados de algumas coleções zoológicas, registrados no *Specify* [Beach, 2010].

No entanto, uma grande parte dos dados sobre as coleções ainda está registrado somente em papel, documentos eletrônicos de texto ou planilhas eletrônicas. Isso significa que as bases de dados eletrônicas não abrangem a totalidade dos registros tombados em todos os acervos institucionais. Os processos de digitalização e verificação dos dados, informações e imagens

[58] http://herbaria.plants.ox.ac.uk/bol/

referente aos registros estão em andamento, fazendo com que a adição de novos dados e correção de informações ocorra com frequência.

Dentre as várias coleções disponíveis, a Coleção de Aves foi a selecionada para a realização de um exemplo de aplicação das tecnologias utilizadas pela WS pelos seguintes motivos:

- O curador da Coleção, Dr. Mário Cohn-Haft, participa de pesquisas voltadas para gerenciamento de dados, metadados e construção de sistemas de informação sobre o domínio de aves e está familiarizado com termos técnicos de análise computacional, o que em alguns aspectos, facilita o diálogo e a compreensão das fases de construção dos sistemas e aplicações;
- A equipe do Dr. Mário (especialistas e estudantes) manifestou extremo interesse e se predispuseram a colaborar com este trabalho de dissertação;
- A coleção de aves, apesar de pequena em número de registros (menor que 10.000 registros), apresenta alto grau de complexidade, permitindo que as soluções para ela desenvolvidas tenham aderência quando reutilizadas para outros domínios de coleções (vertebrados e invertebrados);
- Os dados digitalizados sobre a coleção ainda estão registrados em planilhas, bem como os dados de muitas coleções do instituto, assim representando a necessidade delas no que se refere à coleta, curagem, análise e visualização dos dados.
- A estratégia institucional do PCAC que tem investido na digitalização do acervo e no estabelecimento de mecanismos para integração com iniciativas externas ao INPA (por exemplo, WikiAves[59] e Aves do Amazonas[60]).

[59] http://www.wikiaves.com.br/
[60] http://www.vortexhost.com.br/portais/aves/site/

4.2.
Registrando espécimes sem semântica

O registro das informações associadas ao espécime de uma coleção de dados sobre aves, conforme protocolo padrão, são considerados completos. Cada espécime depositado numa coleção pode conter valiosa informação e quantidade substancial de dados associados. O registro com qualidade e maior quantidade possível de informações que precede a preparação do espécime em si é crucial para que o espécime possa ser aproveitado da melhor maneira possível. Prover informações com clareza e fidelidade certamente tornarão o espécime compreensivelmente útil no acervo e, assim, aumentarão a qualidade e o uso da coleção como um todo.

Segundo o do Manual de Taxidermia[61] da Coleção de Aves do INPA [Macêdo et al., 2011], com o atual questionamento da necessidade de manutenção de coleções biológicas no Brasil e o recorrente debate ético sobre a necessidade de coleta de fauna, o grupo de pesquisa sobre aves do INPA vislumbra a oportunidade de difundir uma metodologia avançada de aproveitamento máximo de espécimes da avifauna coletados em expedições de campo. Por isso, o grupo ressalta o registro de informações com qualidade, recomenda um roteiro básico de dados essenciais e incentiva a coleta de outros itens derivados além da pele, como, por exemplo, imagem, som, tecidos, parasitas, conteúdo estomacal, esqueleto (conforme ilustrado na Figura 8).

Ainda, a qualidade na preparação da pele de uma ave que irá compôr a coleção de peles, por exemplo, também é indispensável, uma vez que algumas características morfológicas da espécie só podem ser avaliadas por meio do exame de um espécime bem preparado. Isso significa que o exemplar deve resguardar as proporções naturais do indivíduo, seguir padrões homogêneos de forma e posição e refletir o aspecto mais natural da plumagem.

[61] "Taxidermia" (termo Grego que significa "dar forma à pele") é a arte de montar ou reproduzir animais para exibição ou estudo. É a técnica de preservação da forma da pele, planos e tamanho dos animais (Hidasi Filho, J., 1976 *apud* http://pt.wikipedia.org/wiki/Taxidermia)

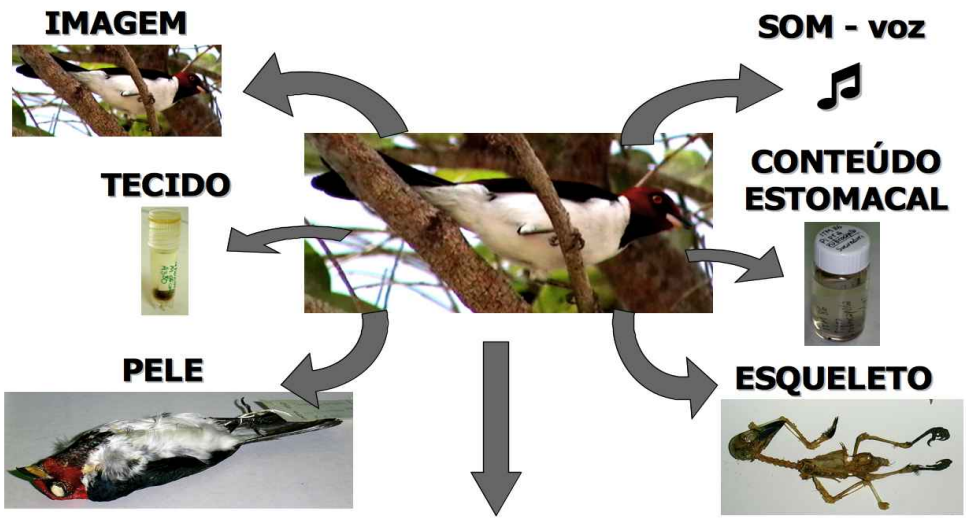


Figura 8 - Itens derivados de um espécime [Macêdo, 2012]

Os dados da Coleção de Aves da Amazônia existentes em meio digital são dados gerais de coleta, taxonomia, preparo do espécime para a coleção de peles e indicativos de presença nas coleções derivadas de tecidos e estômagos. Esses dados são registrados, sem informação semântica, em planilhas eletrônicas que são utilizadas como referência para consultas por espécimes presentes na coleção de peles, localização de coleta, classificação taxonômica e quem são os preparadores dessas peles. Essas consultas são simples e, em alguns casos, como saber se há itens derivados de uma pele específica da coleção de peles na coleção de tecidos, por exemplo, é preciso fazer uma correspondência entre as duas planilhas, a de peles e a de tecidos, para se ter o resultado desejado.

Há casos em que a pele de um espécime de uma determinada espécie não tem correspondente nos registros sobre tecidos, porém, há tecidos daquela mesma espécie que podem corresponder a outros espécimes presentes entre os registros de peles. Isto ocorre porque a curadoria anterior da coleção não adotava a metodologia de aproveitamento máximo da coleção de espécimes da avifauna coletada. E, sendo assim, descartava os itens derivados que poderiam compor as demais coleções.

A recíproca também é verdadeira, há tecidos registrados que não possuem correspondência na coleção de peles pois as circunstâncias da coleta danificaram boa parte da pele inviabilizando a preparação da mesma para compor a coleção.

Além da pele, os itens derivados podem incorporar informações semânticas, consideradas essenciais para a integração de informações com a WS. O modelo

mental (elementos da pesquisa científica e detalhamento das observações) de um pesquisador, poderá ser modelado e associado aos dados e informações da coleção.

4.3. Criação da base de dados semântica ligada

O processo de criação de uma base de dados semântica ligada para o domínio de aves da Amazônia compreende, de acordo com a metodologia utilizada neste trabalho, sete etapas que foram realizadas e são relatadas nas próximas subseções.

4.3.1. Coleta de conjuntos de dados biológicos

Inicialmente, foram realizadas entrevistas com os curadores da coleção de aves objetivando verificar quais os dados sobre aves da Amazônia estão digitalizados, quais os sistemas de informação via Web são utilizados nas consultas pelos usuários em formação na área de Ornitologia[62] e quais são as possíveis integrações com outras bases de dados sobre aves da Amazônia neste momento.

Os curadores de aves disponibilizaram as seguintes planilhas eletrônicas de referência sobre a Coleção de Aves da Amazônia:

- dados sobre a coleção de peles de aves;
- dados sobre a coleção de tecidos de aves;
- lista de espécies para a coleção de tecidos.

Além desses conjuntos de dados disponibilizados, os curadores forneceram, a título de informação semântica sobre o domínio:

- a primeira versão o Manual de Taxidermia de Aves, desenvolvida pelo grupo de pesquisadores ornitólogos do INPA, com informações sobre a taxidermia adotada na preparação das peles de aves;
- um arquivo de apresentação de um curso de diversidade biológica sobre aves.

[62] http://pt.wikipedia.org/wiki/Ornitologia

Os pesquisadores de aves utilizam coleções de livros sobre ornitologia como referência principal. Porém, também consultam informação na Web em sites especializados nessa área de estudo como o do Conselho Brasileiro de Registro Ornitológicos - CBRO[63] e o Portal de Aves do Amazonas[64], inclusive colaborando ativamente com este último. Em visita a esses sites também foram coletados, respectivamente:

- catálogo em planilha eletrônica com dados sobre a taxonomia das aves do Brasil, edição 2009, disponibilizado pelo CBRO[65], adotado pelo o grupo de pesquisa;
- catálogo com dados sobre a taxonomia e nomenclatura das aves do Estado do Amazonas.

Em resumo, o material coletado nesta etapa constitui a base para a realização deste Exemplo de Aplicação.

4.3.2. Modelagem de dados

Os dados registrados na planilha de peles de aves referem-se principalmente à taxonomia (ordem, família, gênero, espécie), à coleta (data, coletor, método de coleta utilizado, localização) e ao preparo (nome, sigla e número do preparador) dos espécimes. Os dados na planilha de tecidos, por sua vez, além de constituírem os mesmos tipos de informação que na de peles, indicam a existência ou não de tecidos de músculo, coração, fígado e sangue dos espécimes coletados, mas não especifica quais são os dados observados sobre esses tecidos. O acervo de conteúdo estomacal, é fisicamente armazenado para identificação e estudo posterior, sendo digitalizado na própria planilha de tecidos apenas uma indicação de existência ou não desse item derivado para cada espécime da coleção.

Considerando-se que não existe um banco de dados para os dados coletados na Etapa 1, mas que existe o CLOSi como esquema de referência, inicialmente foi proposto um modelo com entidades e os seus relacionamentos de modo a realizar uma representação inicial dos dados, conforme a Figura 9. Esta modelagem

[63] http://www.cbro.org.br/CBRO/index.htm
[64] http://www.vortexhost.com.br/portais/aves/site/avesam
[65] http://www.cbro.org.br/CBRO/listabr.htm

conceitual é o ponto de partida aqui adotado para se criar uma base de dados semântica.

No modelo ER proposto, uma entidade biótica está associada a um táxon, que pode ser associado a uma ou mais entidades bióticas. Por outro lado, uma entidade biótica possui uma ou várias localizações, da mesma forma que uma localização possui uma ou mais entidades bióticas. A localização também possui uma ou mais coletas que, por outro lado, possui somente uma localização. Um objeto classificado está associado a uma coleta que pode, por outro lado, ser associada a vários objetos classificados. Um preparador pode preparar vários objetos classificados, bem como um objeto classificado pode ser preparado por vários preparadores. Um objeto classificado pode pertencer a mais de uma coleção e uma coleção deve possuir mais de um objeto classificado. Um objeto classificado é uma entidade biótica que pode ser representada por vários objetos classificados. Da mesma forma que uma coleção pode ser de peles ou de tecidos, mas cada tipo é uma coleção distinta.

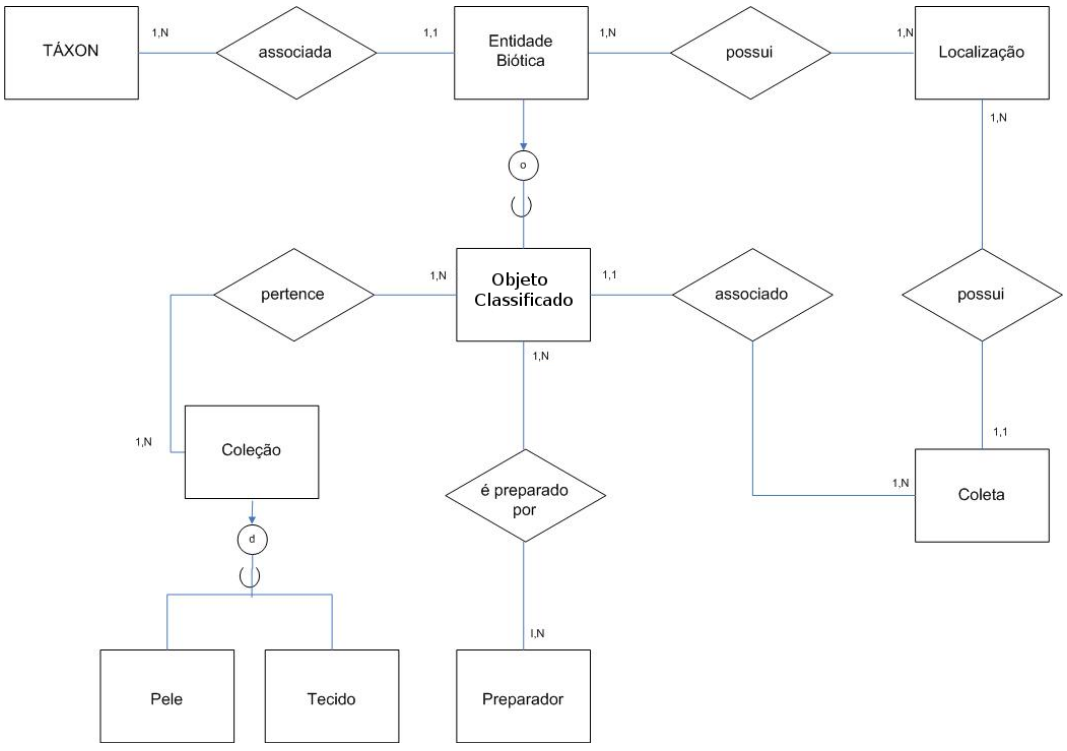


Figura 9 - Modelo ER para a Coleção de Aves da Amazônia do INPA

Ainda, o modelo ER proposto possui correspondência com os *clusters* do CLOSi. A entidade Taxon está contida no escopo do *cluster Taxonomy*. As entidades Objeto_Classificado e Coleção estão contidas no escopo do *cluster*

Collection_Management. A entidade Localização está contida no escopo do *cluster Locality_Of_Biodiversity_Data.* A entidade Coleta está contida no escopo do *cluster Collecting_Event_Of_Collection*. E, por fim, a entidade Preparador está contida no escopo do *cluster Agent_Of_Collection*. Portanto, o modelo é altamente aderente ao CLOSi.

Uma vez definido um modelo conceitual, foi necessária a criação de um modelo lógico para a implementação do banco de dados. Neste modelo podem ser evidenciados os atributos e seus tipos para cada tabela do banco de dados. Ainda, as relações N:N são representadas como tabelas e também podem conter atributos. A Figura 10 apresenta a implementação do modelo lógico criado para os dados digitalizados da Coleção de Aves da Amazônia do INPA.

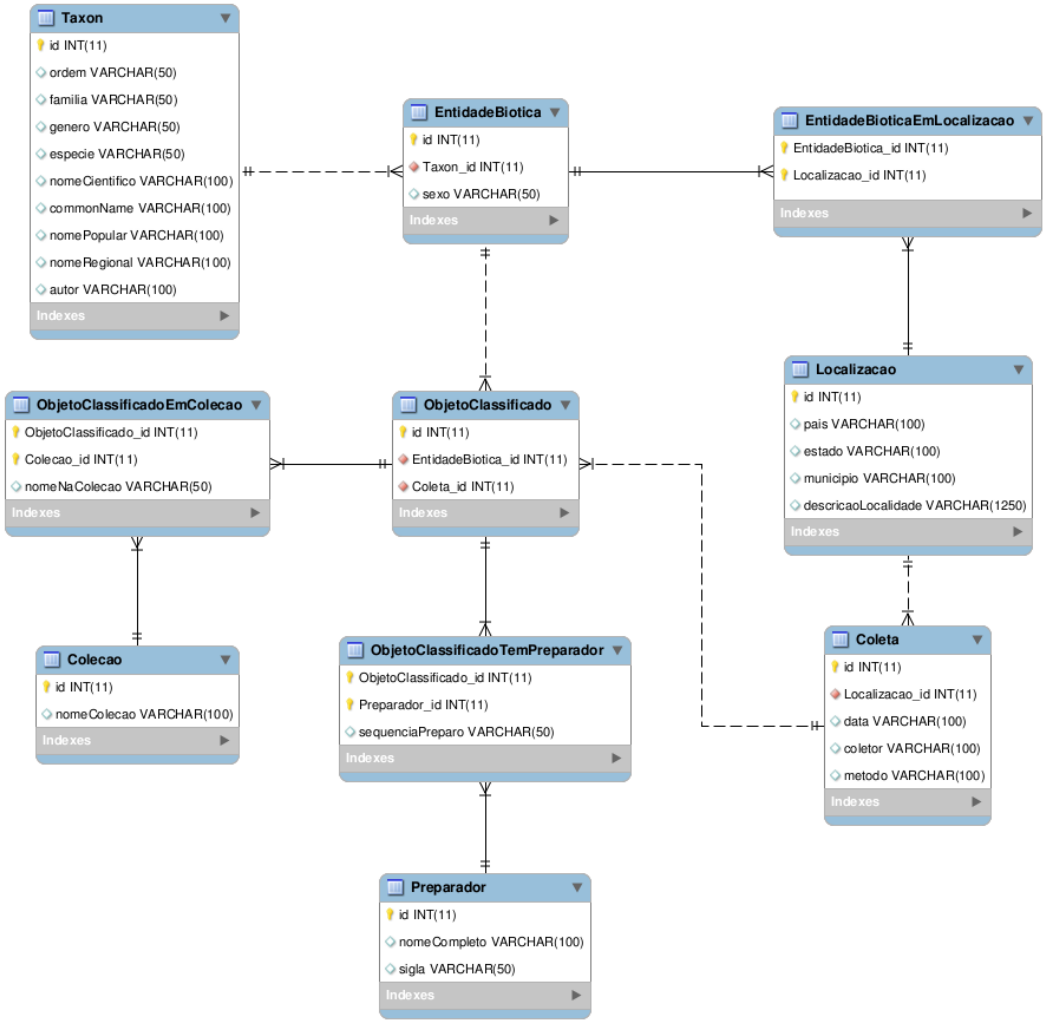


Figura 10 - Modelo lógico do banco de dados da Coleção de Aves da Amazônia do INPA

Uma vez criada a estrutura do banco de dados, era necessária a alimentação das tabelas através da etapa de extração e transformação dos dados das planilhas eletrônicas.

4.3.3.
Seleção e extração de dados sobre aves

No processo de construção de uma base de dados semântica a partir de conjuntos de dados legados, uma etapa importante a ser realizada é a seleção e extração dos dados, geralmente em fontes mal estruturadas, para que seja realizado um pré-processamento destes dados e/ou seu armazenamento.

Para a composição da bases de dados semântica de aves da Amazônia, foi realizada a organização dos dados das planilhas disponibilizadas pelo grupo de pesquisa de aves do INPA, bem como as planilhas coletadas nos sites do CBRO e do Portal de Aves do Amazonas:

- Da planilha de dados sobre a coleção de peles, foram selecionados e extraídos os dados de taxonomia (campos: ORDEM, FAMÍLIA, GÊNERO, ESPÉCIE), de localização (campos: PAIS, ESTADO, MUNICÍPIO e LOCALIDADE), da coleta (campos: COLETOR, DATA_COLETA e MET_COLETA), do preparador (campos: PREPARADOR, Sigla Prep e Seq. Prep.), da coleção (campo: n° INPA) e da entidade biótica em si (campo: SEXO);
- Da planilha de dados sobre a coleção de tecidos, foram selecionados e extraídos os dados de taxonomia (campos: GENERO e ESPECIE), de localização (campos: EST e LOCALIDADE), da coleta (campos: MÉTODO DE COLETA, DATA_COLETA), da coleção (campo: No TEC.) e da entidade biótica em si (campo: SEXO);
- Da planilha com a lista de espécies para tecidos e do site do CBRO, foram selecionados e extraídos os dados de taxonomia (campos: Ordem, Família, CBRO 2009 Gênero, CBRO 2009 Espécie, Autor);
- Da planilha com o catálogo de Aves do Amazonas, foram selecionados e extraídos os dados de taxonomia (campos: Nome Científico, Common name, Nome popular, Nome regional).

A seleção e extração dos dados dessas planilhas foi implementada com a ferramenta Kettle, conforme mencionado na secção 3.2.3, bem como o armazenamento do resultado do processo de extração e transformação dos dados na base relacional criada na Etapa 2.

4.3.4.
Curagem dos dados digitais

A curadoria de dados compreende uma gama de atividades e processos realizados para criar, gerenciar, manter e validar um conjunto de dados de pesquisa com qualidade, permitindo que possam estar disponíveis para reutilização e preservação por longos períodos. Na ciência, curadoria de dados pode indicar um processo de extração de informações importantes de textos científicos, tais como artigos e relatórios de pesquisas elaborados por especialistas, onde as informações serão convertidas em formato eletrônico, para uma possível entrada em um banco de dados biológicos.

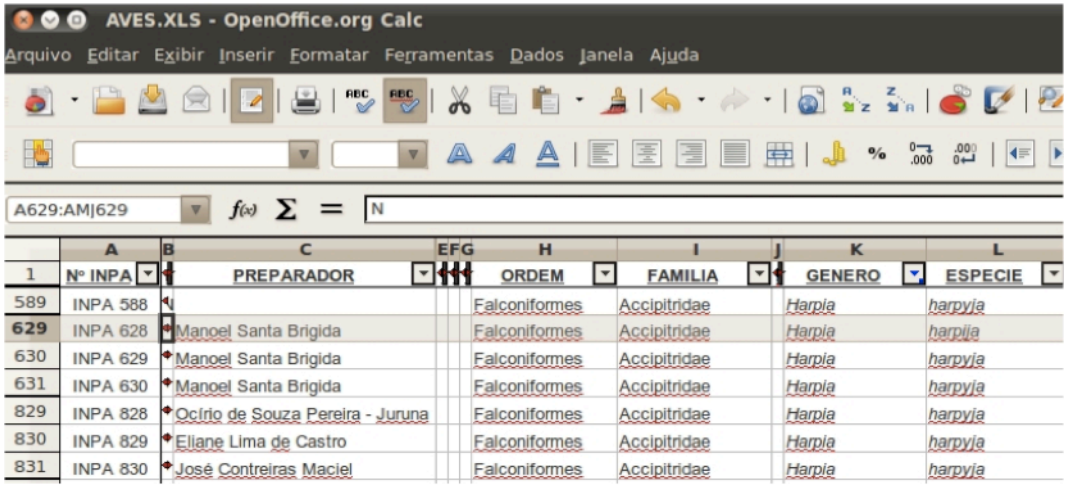
No ambiente de coleções biológicas erros ocorrem independentemente do delineamento de experimentos, de sua condução, e implementação de estratégias de prevenção de erros. O processo de limpeza de dados objetiva identificar e corrigir esses erros ou pelo menos minimizar seus impactos nos resultados de estudos. Em geral, pouca atenção é dispensada no tema e também, existe pouca orientação disponível sobre como organizar e conduzir a limpeza de forma eficiente e ética.

Um desenvolvimento esperado com bastante ênfase está no campo da padronização, documentação e comunicação de manipulação e qualidade de dados. Na tradição acadêmica a validade de estudos tem sido discutida predominantemente com relação ao delineamento, atendimento de protocolo geral e a integridade e experiência do pesquisador. A manipulação de dados, embora tenha a capacidade de afetar a qualidade dos resultados das pesquisas, tem recebido menor atenção. Com isso, existem lacunas de conhecimento sobre uma metodologia ótima para a manipulação de dados e padrões de qualidade de dados.

Neste trabalho considera-se a limpeza de dados como processo em três etapas, envolvendo repetidos ciclos de triagem, diagnóstico e edição de dados com suspeitas de anormalidades. É reconhecido que muitos erros de dados são detectados incidentalmente durante o estudo e manipulação trivial e não durante o processo de limpeza dos dados. Entretanto, é mais eficiente a detecção pela busca ativa e intencional, de forma planejada.

A curagem dos dados pôde ser implementada através do *Google Refine*, com a identificação, diagnóstico e correção dos dados. Como pode ser observado

na Figura 11, o nome da espécie *harpyja* do item INPA 628 da coleção de peles foi digitado erroneamente com "i" ao invés de "y". Erros comuns como esse são facilmente identificados e corrigidos com o Google Refine.



	A	B	C	EFG	H	I	J	K	L
1	Nº INPA		PREPARADOR		ORDEM	FAMILIA		GENERO	ESPECIE
589	INPA 588				Falconiformes	Accipitridae		Harpia	harpyja
629	INPA 628		Manoel Santa Brigida		Falconiformes	Accipitridae		Harpia	harpija
630	INPA 629		Manoel Santa Brigida		Falconiformes	Accipitridae		Harpia	harpyja
631	INPA 630		Manoel Santa Brigida		Falconiformes	Accipitridae		Harpia	harpyja
829	INPA 828		Ocírio de Souza Pereira - Juruna		Falconiformes	Accipitridae		Harpia	harpyja
830	INPA 829		Eliane Lima de Castro		Falconiformes	Accipitridae		Harpia	harpyja
831	INPA 830		José Contreiras Maciel		Falconiformes	Accipitridae		Harpia	harpyja

Figura 11 - Exemplo de erro de registro de dados

A Figura 12 ilustra a identificação, edição e correção do erro através da percepção de que existem seis itens com o nome da espécie digitado da forma correta, *harpyja*, e uma forma errada, *harpija*. Um erro como esses, em uma planilha com duas mil e cem entradas é dificilmente perceptível ao olho humano, mas rapidamente identificado por uma ferramenta de limpeza de dados.

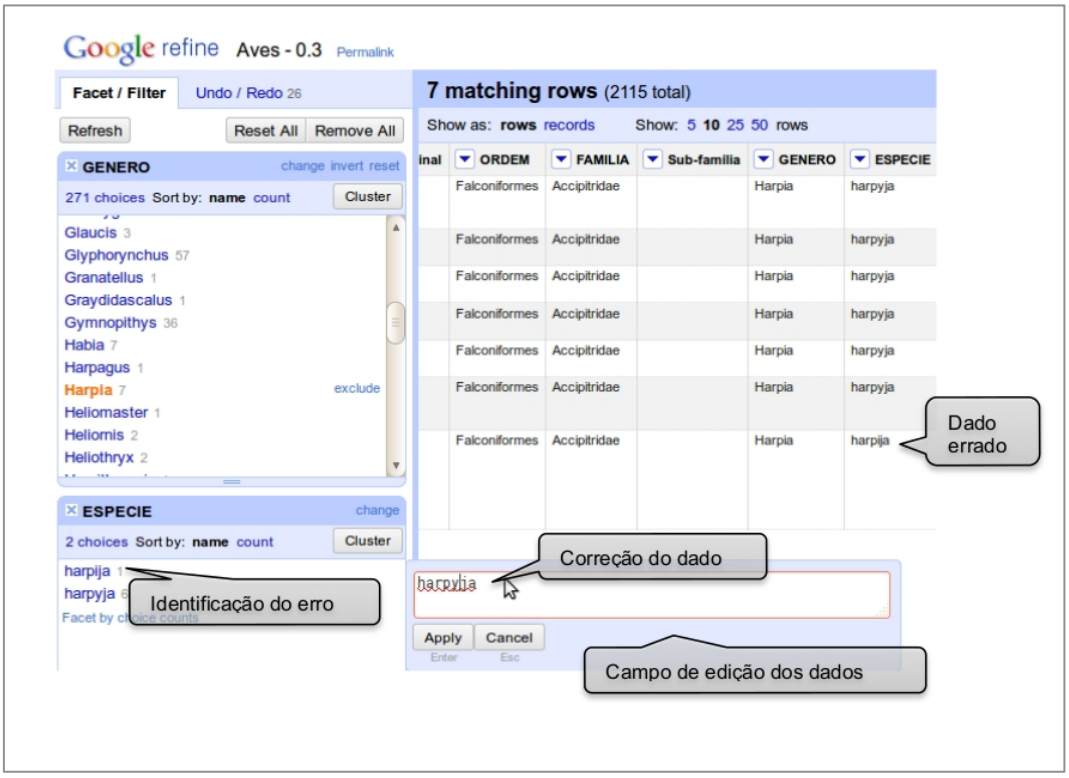


Figura 12 - Identificação, edição e correção de erro utilizando o Google Refine

Um resumo com a curagem dos dados digitais das coleções de peles e tecidos de aves do INPA pode ser verificada na Tabela 1.

Tipos de Curagem	Quantidade de Edições
Normalização de nomes dos preparadores	1 (114 registros alterados)
Normalização de siglas dos preparadores	6 (185 registros alterados)
Normalização de nomes dos coletores	6 (27 registros alterados)
Normalização de nomes dos coletores nas etiquetas dos itens de coleção	9 (28 registros alterados)
Correção de nomenclatura de “Gênero”	7 (53 registros alterados)
Correção de nomenclatura de “Espécie”	2 (2 registros alterados)
Correção de nomenclatura de “Sub -família”	1 (3 registros alterados)

Tabela 1 - Resumo da curagem dos dados digitais da Coleção de Aves do INPA

4.3.5. Mapeamento para base de dados triplificada

O mapeamento de uma base de dados relacional para uma base de dados triplificada envolve a conversão dos dados para uma base serializada em RDF. Esta serialização é alcançada com a utilização de um vocabulário ou ontologia que define a representação daqueles dados em um modelo baseado em triplas de dados.

A ontologia adiciona informações sobre os dados, ou seja, agrega semântica à base de dados durante a etapa de mapeamento. Considerando-se o estudo realizado por Albuquerque para a construção da OntoBio, foi proposta uma ontologia, *Amazonian Birds Collection Ontology*[66] - ABC (Figura 16) para a realização do Exemplo de Aplicação.

As classes e propriedades da ABC são baseadas nas classes e propriedades correspondentes da OntoBio. Algumas classes da OntoBio foram suprimidas visando mapear a base de dados relacional criada para uma base de dados semântica utilizando-se uma ontologia mais aderente ao modelo conceitual concebido na Etapa 2, porém garantindo a expressividade semântica.

[66] http://lis.inpa.gov.br/ontology/abc

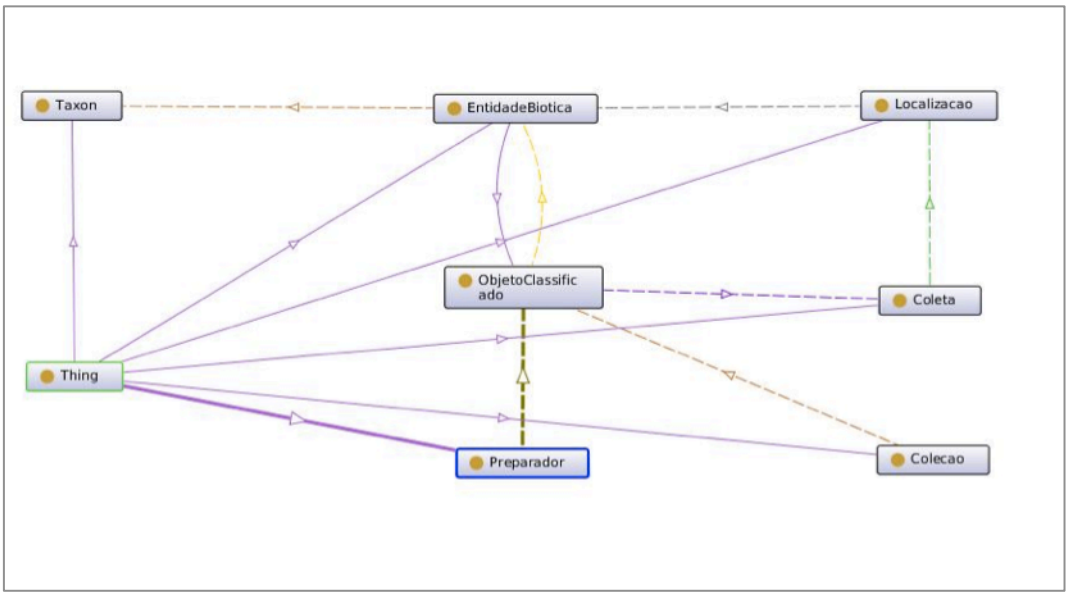


Figura 13 - *Amazonian Birds Collection Ontology* (ABC)

As classes da ABC são: Taxon, EntidadeBiotica, Localizacao, ObjetoClassificado, Coleta, Preparador e Coleta. A agregação semântica ocorre justamente quando as anotações são feitas na composição de triplas.

Por exemplo, o Quadro 5 apresenta triplas para o Taxon 1, formadas no mapeamento para a base de dados semântica utilizando-se a ABC.

```
<http://localhost:8080/lis/resource/Taxon/1>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://localhost:8080/ontology/Taxon> .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temOrdem> "Falconiformes" .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temFamilia> "Accipitridae" .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temGenero> "Harpia" .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temEspecie> "harpyja" .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temNomeCientifico> "Harpia harpyja" .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temCommonName> "Harpy Eagle" .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temNomePopular> "gavião-real" .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temNomeRegional> "Não Informado" .

<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/temNomeAutor> "Linnaeus, 1758" .
```

Quadro 5 - Serialização N-Triples para *Harpia harpyja*

As propriedades de dados, como temOrdem, temFamília, temGenero, temEspecie, agregam mais significado que os atributos **ordem**, **familia**, **genero** e **especie**, respectivamente, do modelo conceitual. Por exemplo, enquanto no modelo conceitual "Taxon 1" é uma instância da entidade **Taxon** que possui o atributo **ordem** igual ao valor "Falconiformes", na ABC <Taxon 1> é um recurso que possui um relacionamento <temOrdem>, definido pela ontologia, com o valor literal <Falconiformes>. De acordo com a semântica que se queira expressar, <temOrdem> pode relacionar <Taxon 1> a outro recurso como, por exemplo, o <http://dbpedia.org/resource/Falconiformes>. Resumindo, <temOrdem> pode relacionar recursos ao invés de somente atribuir valores.

As propriedades de objetos, como eClassificacaoDe e ocorreNaLocalizacao, apresentadas no Quadro 6, agregam mais significado que os relacionamentos **associada** e **possui** do modelo conceitual. Por exemplo, enquanto no modelo conceitual "Entidade Biotica 1" é uma instância da entidade **Entidade Biotica** e possui o relacionamento **possui** com a instância "Localizacao 1" da entidade **Localizacao**, na ABC <EntidadeBiotica 1> é um recurso que possui um relacionamento <ocorreNaLocalizacao>, definido pela ontologia, com o recurso <Localizacao 1>. A definição do relacionamento na ABC expressa de forma mais clara não o sentido de "posse", mas que as entidades bióticas ocorrem nas localizações onde aconteceram as coletas.

```
<http://localhost:8080/lis/resource/Taxon/1>
<http://localhost:8080/ontology/eClassificacaoDe>
<http://localhost:8080/lis/resource/EntidadeBiotica/3> .

<http://localhost:8080/lis/resource/EntidadeBiotica/1>
<http://localhost/ontology/ocorreNaLocalizacao>
<http://localhost:8080/lis/resource/Localizacao/1> .
```

Quadro 6 - Exemplos de propriedades de objetos

4.3.6. Ligação da base de dados triplificada com a LOD

Após a fase de mapeamento ou triplificação, as práticas *Linked Data* recomendam que sejam realizadas ligações da base criada com outras bases da nuvem LOD, fornecendo, assim, contexto.

O levantamento de potenciais alvos de ligação com a base de dados semântica de aves considerou a proveniência dos dados com os quais a base seria

ligada e a utilidade dessas ligações para os pesquisadores e alunos de pós-graduação de ornitologia.

Os dados sobre aves ainda são escassos na LOD e a principal fonte de proveniência em se tratando de dados sobre aves da Amazônia é a Wikipedia. Existem informações também disponibilizadas pelo portal da BBC *Nature*, do Projeto *TaxonConcept*[67], nas bases de conhecimentos *Geospecies*[68] e *Freebase*[69].

O curador da Coleção de Aves do INPA, Dr. Mario Cohn-haft, considerou que as potenciais fontes de ligação são válidas e que, apesar de não terem o rigor científico, são úteis para fins de colaboração uma vez que permitem a atualização de conteúdo como a correção de informação incorreta ou a complementação de informações incompletas.

A verificação de similaridade de recursos pode ser realizada com as ferramentas Silk e LIMES. Uma vez comprovada a equivalência entre dois recursos, é realizada a ligação.

No processo de ligação de dados, foram criados links de relacionamento para as propriedades de dados de taxonomia (temOrdem, temFamilia, temGenero) e de localização (estaNoPais, estaNoEstado, estaNoMunicipio); e links de identidade, principalmente através do tipo owl:sameAs.

O Quadro 7 mostra as ligações entre o recurso *Harpia harpyja* da base criada neste trabalho e os recursos equivalentes das bases ligadas da *Dbpedia*[70], da *BBC Nature*, do *TaxonConcept*, da *Geospecies* e da *Freebase*.

```
<http://localhost:8080/lis/resource/Taxon/1>
<http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Harpy_Eagle> .

<http://localhost:8080/lis/resource/Taxon/1>
<http://www.w3.org/2002/07/owl#sameAs>
<http://www.bbc.co.uk/nature/life/Harpy_Eagle#species> .

<http://localhost:8080/lis/resource/Taxon/1>
<http://www.w3.org/2002/07/owl#sameAs>
<http://lod.taxonconcept.org/ses/DzFEQ#Species> .

<http://localhost:8080/lis/resource/Taxon/1>
<http://www.w3.org/2002/07/owl#sameAs>
<http://lod.geospecies.org/ses/DzFEQ> .

<http://localhost:8080/lis/resource/Taxon/1>
<http://www.w3.org/2002/07/owl#sameAs>
<http://rdf.freebase.com/ns/m.01_36n> .
```

Quadro 7 - Ligação de recursos para *Harpia harpyja*

[67] http://www.taxonconcept.org/
[68] http://lod.geospecies.org/
[69] http://www.freebase.com/
[70] http://dbpedia.org/About

O processo de descoberta de relações entre as entidades da base de dados triplificada local e entidades localizadas em outros bancos de triplas não é trivial e envolve a participação dos especialistas no domínio visando a validade das ligações a serem consolidadas.

No que se refere aos links de relacionamento, houve a facilidade de se encontrar correspondência para os dados das categorias taxonômicas de ordem e família e para os dados de localização de país e estado, por serem mais genéricos; e dificuldade de se encontrar correspondência para os dados da categoria taxonômica de gênero e para os dados de localização de município, por serem mais especializados. Sendo assim, não puderam ser estabelecidos links de relacionamento para todos os gêneros e nem para todos os municípios a base triplificada local.

No que se refere aos links de identidade, houve facilidade na verificação de similaridade para os dados encontrados na bases da *DBpedia*, do *TaxonConcept* e da *Freebase*; e dificuldade de se checar a similaridade nas bases *BBC* e *Geospecies*, uma vez que os SPARQL *Endpoints* dessas bases estavam indisponíveis a maior parte do tempo.

4.3.7. Armazenamento da base de dados triplificada

A etapa final da metodologia para a criação de uma base de dados semântica ligada é o seu armazenamento em bancos de triplas. Neste trabalho, ao final do processo ETL com a ferramenta Kettle, é gerado um arquivo texto com as triplas resultantes do processo. Este arquivo é carregado na ferramenta *OpenRDF Workbench* onde são gerados os arquivos a serem carregados no repositório Sesame no Rexplorator.

A base criada para a prova de conceito deste trabalho utilizou 7 classes, 35 propriedades e foi gerada com aproximadamente 670 triplas.

Uma vez a base de dados semântica criada com as informações sobre a coleção de aves da Amazônia, o pré-requisito para a manipulação de dados ligados está atendido e podem ser construídas consultas e aplicações Web para o domínio em questão.

4.4. Exploração da base RDF com o Rexplorator

Esta seção aborda a construção de consultas com o Rexplorator baseadas no modelo RDFS da base de dados semântica ligada criada para a Coleção de Aves da Amazônia.

Para tanto, foram elaborados alguns casos de uso com a participação dos curadores da coleção que permitem ao usuário explorar a base semântica criada. A seguir estes casos de uso são descritos para a compreensão da abordagem do problema com a utilização da ferramenta.

4.4.1. Casos de uso

1. **Consultar espécimes por entidade biótica**. Neste caso de uso é exibida uma lista com as entidades bióticas presentes na coleção. Ao selecionar uma das opções, seriam exibidas os espécimes presentes na base de acordo com a opção escolhida. Neste momento, o usuário pode selecionar um dos espécimes listados para consultar mais dados disponíveis na base sobre o mesmo.
2. **Consultar espécimes por categoria taxonômica**. Neste caso de uso é exibida uma lista com as opções de consulta gênero e espécie, o que significa o nome científico da ave presente na coleção. Ao selecionar uma das opções, são exibidas as entidades bióticas presentes na base de acordo com a categoria taxonômica escolhida. Neste momento, o usuário pode selecionar um dos itens listados para consultar quais os espécimes que representam aquela entidade biótica classificada com aquele táxon. Ainda, o usuário pode consultar mais dados disponíveis na base sobre os espécimes que aparecerem como resultado da consulta.
3. **Consultar espécimes por coleção**. Neste caso de uso é exibida uma lista com as opções de coleção existentes (peles, tecidos ou estômagos). Ao selecionar uma das opções, é exibida uma lista com os espécimes de aves presente na coleção escolhida. Neste momento, o usuário pode selecionar um dos espécimes listados para consultar

mais dados disponíveis sobre o mesmo na base, inclusive a correspondência entre as coleções, por exemplo, se existem itens derivados daquele espécime em outras coleções.

4. **Consultar espécimes por preparador**. Neste caso de uso é exibida uma lista com os preparadores dos espécimes que estão presentes na Coleção de Aves do INPA. Ao selecionar um preparador, é exibida uma lista com os espécimes preparados por ele. Neste momento, o usuário pode selecionar um dos espécimes listados para consultar mais dados disponíveis sobre o mesmo na base.
5. **Consultar espécimes por coleta**. Neste caso de uso é exibida uma lista com os locais de coleta utilizados para captura dos espécimes que estão presentes na Coleção de Aves do INPA. Ao selecionar um local, é exibida uma lista com os espécimes coletados naquele sítio específico. Neste momento, o usuário pode selecionar um dos espécimes listados para consultar mais dados disponíveis sobre o mesmo na base.
6. **Consultar espécimes por localização**. Neste caso de uso é exibida uma lista com as opções de locais de coleta utilizados nas coletas dos espécimes da Coleção de Aves da Amazônia do INPA. Ao selecionar uma das opções, é exibida uma lista com os itens presentes na coleção que foram coletados de acordo com a localização escolhida. Neste momento, o usuário pode selecionar um dos espécimes listados para consultar mais dados disponíveis sobre o mesmo na base.
7. **Consultar um nome científico**. Neste caso de uso é exibido para o usuário um campo texto para que ele digite o nome científico sobre o qual quer informações. Ao submeter o formulário, são exibidas informações sobre o nome científico procurado. Essas informações podem conter tanto dados presentes na base local quanto dados recuperados pelas ligações com outros recursos na LOD.

4.4.2. Implementação da aplicação Web

A aplicação Web foi implementada utilizando o compartilhamento fechado de consultas no ambiente de autoria do Rexplorator. Esse tipo de aplicação pode ser utilizado por usuários que não tenham conhecimento algum em RDF ou Web Semântica, sendo possível trazer parte dos benefícios desta para usuários comuns com conhecimentos básicos em computação.

A seguir são demonstrados detalhes de implementação de casos de uso da aplicação para o entendimento do processo de construção de consultas com a ferramenta. Em [Azevedo, 2010] são apresentados mais detalhes de implementação e todos os recursos da ferramenta.

O primeiro passo para se gerar o primeiro caso de uso "espécimes por entidade biótica" é montar em um *workbench* uma consulta que retorne os espécimes da coleção de aves que representam uma entidade biótica. Esse será o *workbench* que irá conter o primeiro caso de uso. Para isso, o usuário que está montando aplicação deve realizar uma consulta que retorne todas as entidades bióticas da base. Isto é possível fixando a propriedade "eRepresentadaPor" na posição de propriedade e realizando uma consulta SPO, através da qual se obtém um conjunto com todas as entidades bióticas da base que são representadas por algum espécime (Figura 14).

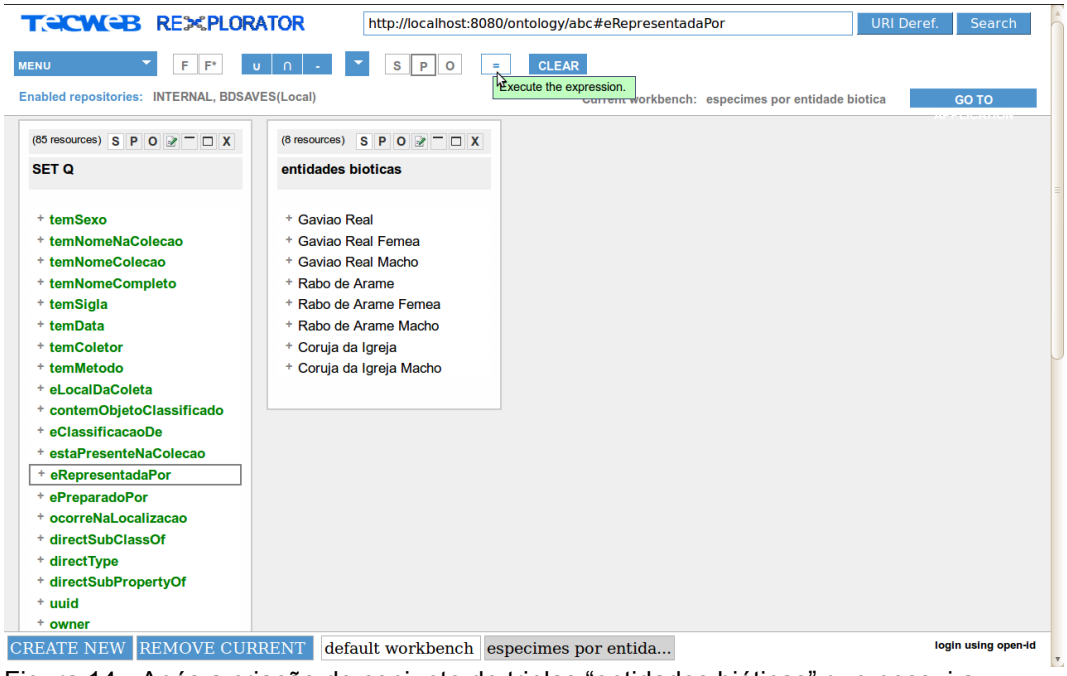


Figura 14 - Após a criação do conjunto de triplas "entidades bióticas" que possui a propriedade "eRepresentadaPor" na posição de predicado

Nesse momento, é possível realizar uma nova consulta SPO, selecionando uma entidade biótica na posição de sujeito e a propriedade "eRepresentadaPor" na posição de propriedade. A consulta realizada retorna os espécimes que representam a entidade biótica selecionada (Figura 15).

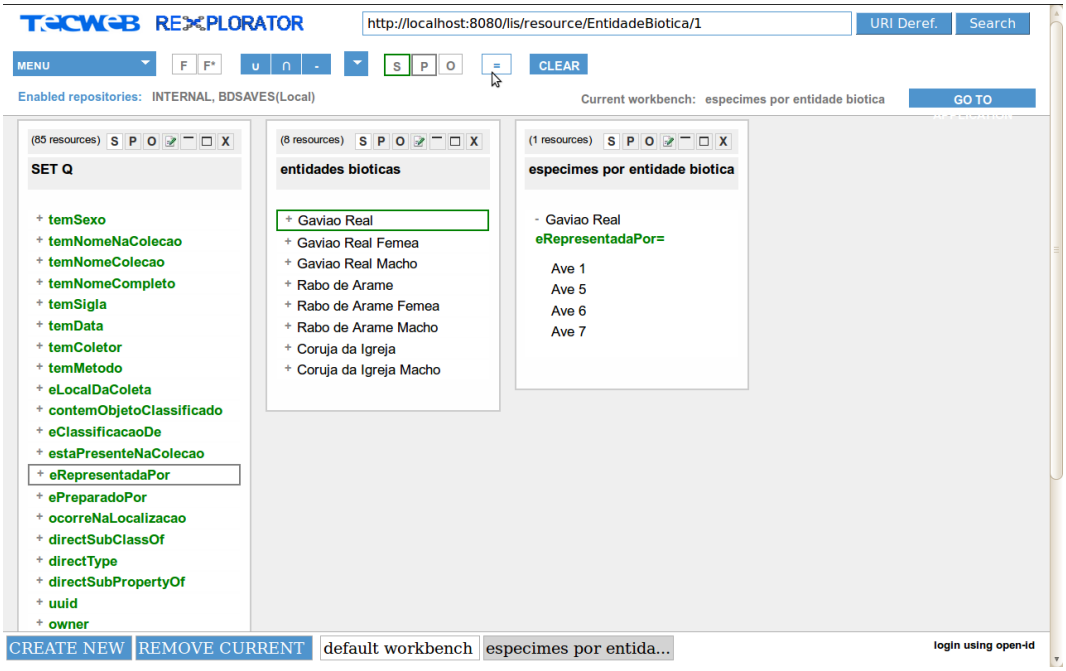


Figura 15 - Criação do conjunto "espécimes por entidade biótica"

Para facilitar o entendimento, os conjuntos gerados foram renomeados com nomes que representam as semânticas dos conjuntos. O primeiro recebeu o nome "entidades bióticas" e o segundo recebeu o nome "espécimes por entidade biótica". O *workbench*, por sua vez, foi renomeado para "espécimes por entidade biótica" que será o nome que aparecerá no menu principal da aplicação Web.

Para continuar a geração da aplicação, o usuário deve acessar o modo de edição dos conjuntos gerados. A primeira tarefa do modo de edição é parametrizar a entidade biótica selecionada no segundo conjunto (Figura 16). Dessa forma, será possível reutilizar a consulta com outras entidades bióticas.

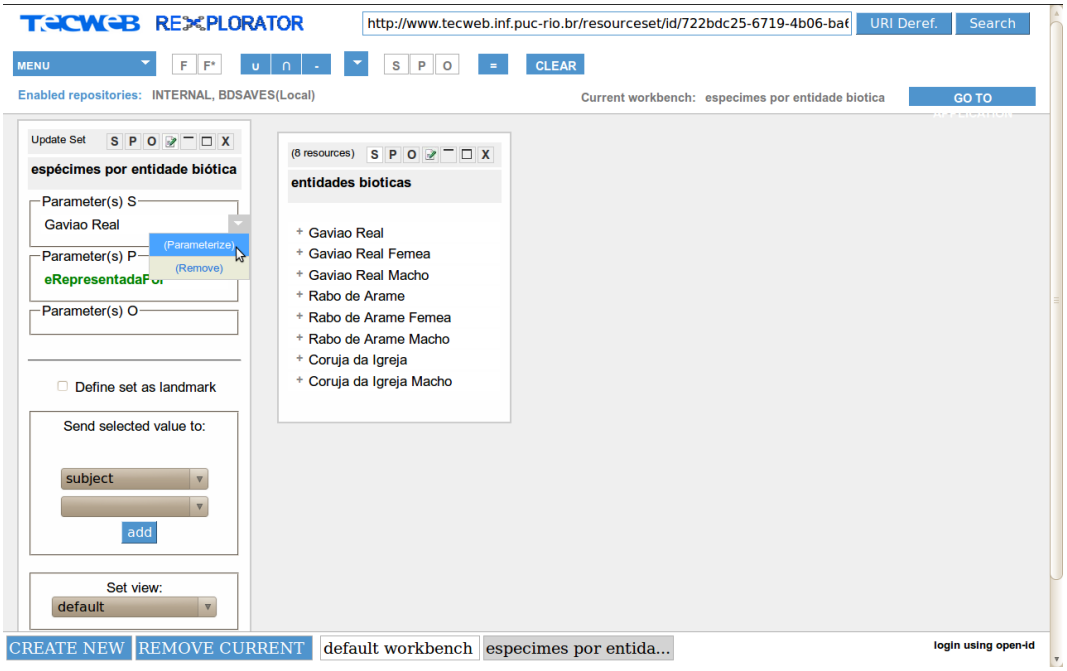


Figura 16 - Entidade biótica dos espécimes sendo parametrizada para tornar possível a reutilização da consulta

Uma vez que a consulta está parametrizada, deve-se definir os conjuntos que devem ser exibidos no momento em que a consulta for acessada. A aplicação deve apresentar um ponto de partida para que o usuário final que a estiver acessando possa alimentar os parâmetros da consulta montada. Neste caso de uso, o ideal é que o primeiro conjunto montado, uma lista com todas as entidades bióticas existentes na base, seja o ponto de partida, apresentando as opções disponíveis para o parâmetro da consulta. Para marcar o conjunto como ponto de partida para o caso de uso, basta selecionar a opção "*Define set as landmark*" no seu modo de edição (Figura 17). Isso significa que os recursos do conjunto serão exibidos como opções quando o caso de uso em questão for selecionado do menu da aplicação Web gerada.

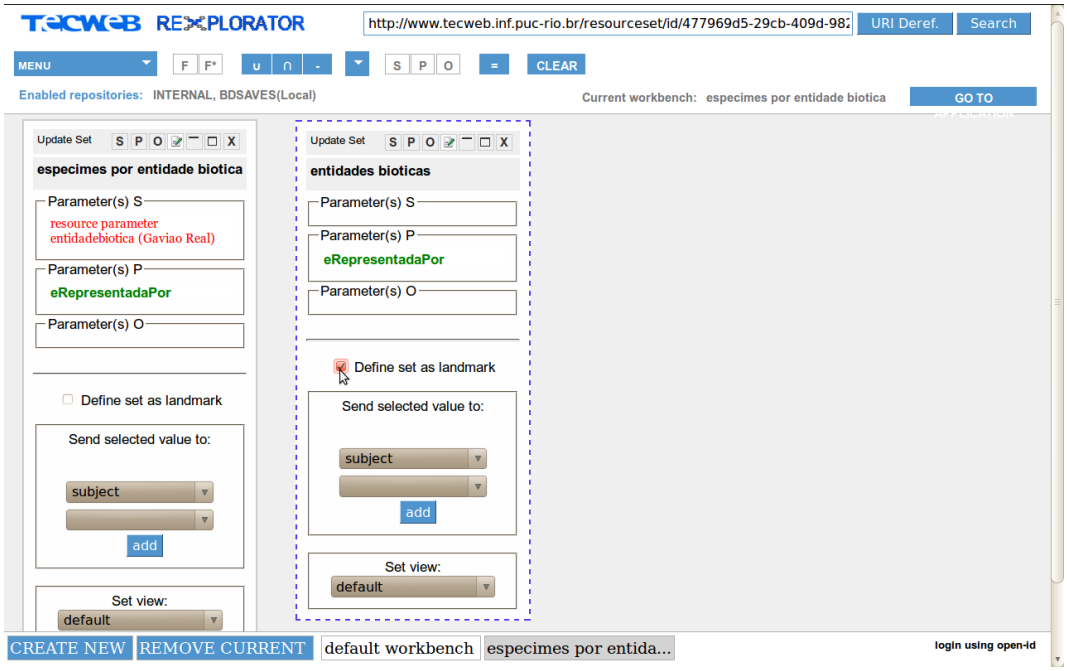


Figura 17 - Definição do conjunto de triplas como ponto de partida para o caso de uso

O segundo passo é dar a semântica para a ação do usuário final de seleção de uma entidade biótica. Quando o usuário final selecionar uma entidade da lista exibida, o segundo conjunto deve ser gerado com a entidade selecionada substituindo o valor do parâmetro e os resultados obtidos devem ser exibidos. Esse índice de entidades bióticas é, de fato, um conjunto de triplas, com as entidades bióticas na posição do sujeito. No caso do exemplo, o comportamento de escolha deve ser associado somente ao índice formado pelos sujeitos das triplas do conjunto de triplas. Para adicionar esse comportamento ao conjunto que possui as entidades, deve-se selecionar, no primeiro *combo-box* de comportamento, o índice no qual o comportamento está sendo adicionado (no caso do exemplo, é o índice formado pelos recursos que estão na posição do sujeito das triplas). Depois, deve-se selecionar, no segundo *combo-box* de comportamento, que exibe todos os parâmetros de conjuntos de triplas existentes na aplicação, o conjunto que será avaliado e exibido e o parâmetro que o valor selecionado pelo usuário final deve substituir (Figura 18). A posição da tripla escolhida no primeiro *combo-box* e o nome do parâmetro e conjunto escolhidos serão exibidos em uma lista, confirmando o comportamento desejado (Figura 19).

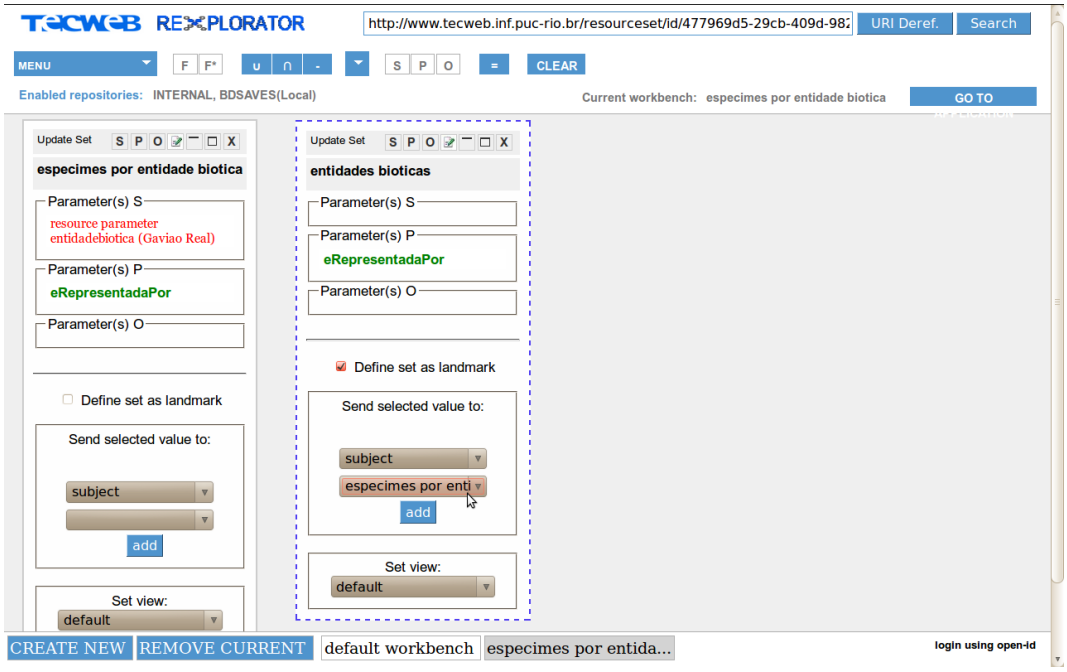


Figura 18 - Adicionando o comportamento de passagem de valor para parâmetro ao conjunto de triplas

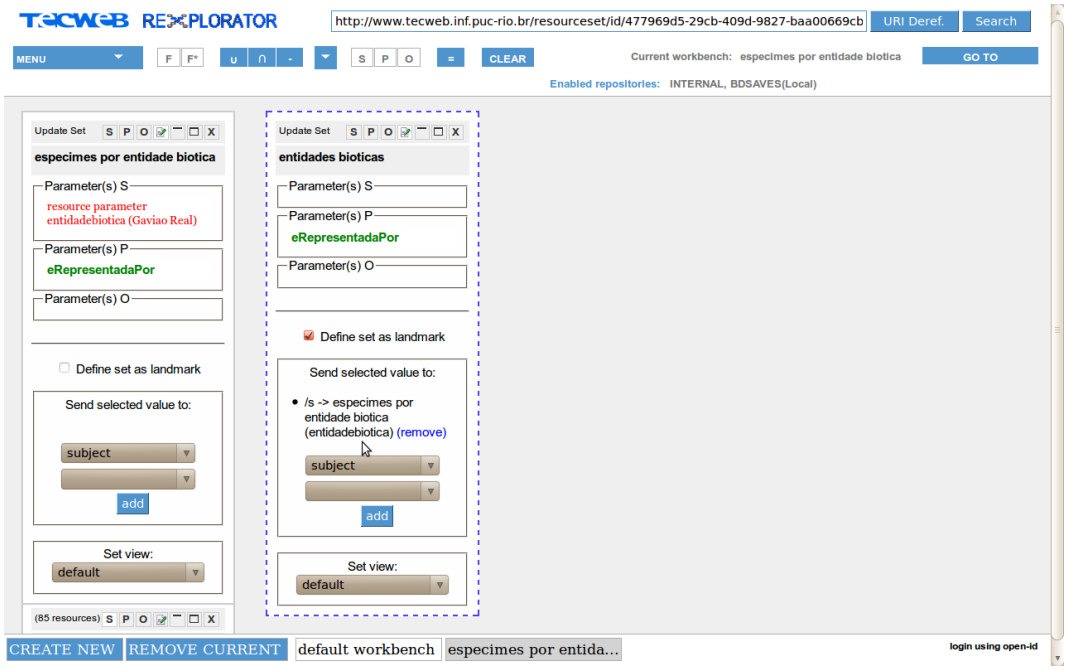


Figura 19 - Indicação que o valor do recurso na posição do sujeito que for selecionado será usado como parâmetro do outro conjunto de triplas

Com os comportamentos definidos, quando o usuário final da aplicação gerada acessar através do menu o caso de uso “espécimes por entidade biótica”, será exibida para ele a listagem com todos os recursos do conjunto das entidades bióticas disponíveis. Uma vez que ele selecione alguma delas, as triplas do segundo conjunto serão calculadas utilizando a entidade selecionada no lugar do parâmetro e o resultado será exibido ao usuário final.

Com o objetivo de tornar o caso de uso mais completo, é interessante que o usuário final possa selecionar algum dos espécimes da lista de espécimes de uma entidade biótica previamente selecionada e consiga visualizar informações adicionais sobre o mesmo disponíveis na base, como qual o seu nome na coleção, em quais coleções está presente e quem preparou aquele espécime para a coleção. Para isso é necessário criar um novo conjunto de triplas, que possua uma consulta SPO com essas propriedades na posição P e a posição S parametrizada. O espécime selecionado pelo usuário final será utilizado como valor da posição S. Uma boa forma de acessar as propriedades necessárias é visualizando todas as propriedades de um espécime (Figura 20).

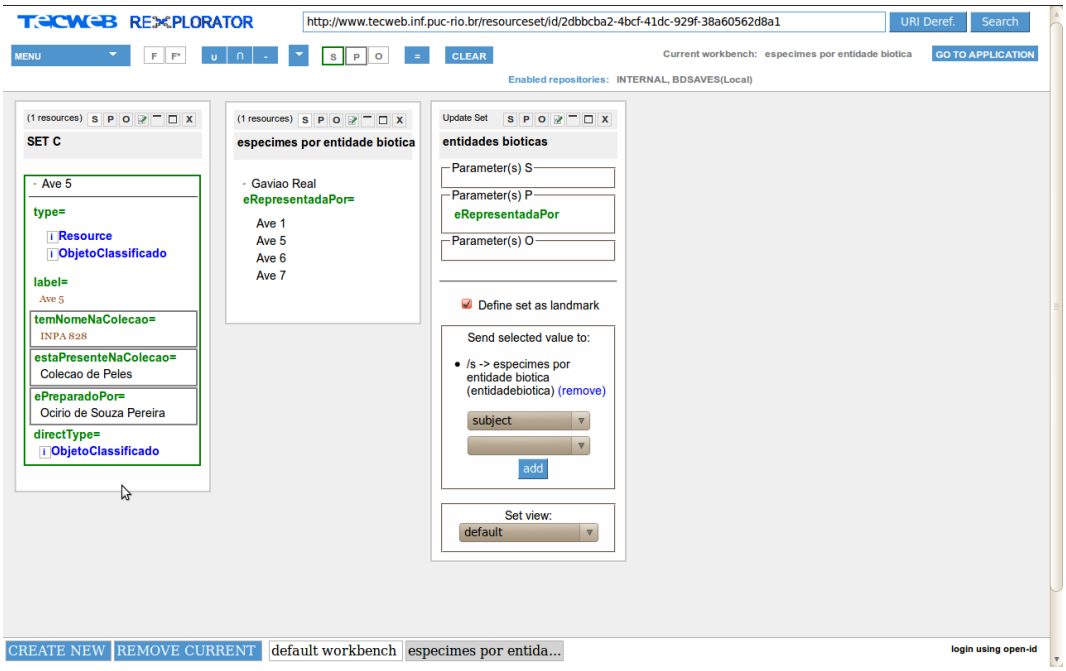


Figura 20 - Criação de um conjunto com as propriedades desejadas através de uma operação SPO

Uma vez que o conjunto foi criado utilizando um espécime qualquer na posição S, é necessário parametrizá-lo através do modo de edição do conjunto. Também é uma boa prática renomear o conjunto com um nome que represente a sua semântica. Nesse caso, “espécime” pode ser utilizado como nome (Figura 21).

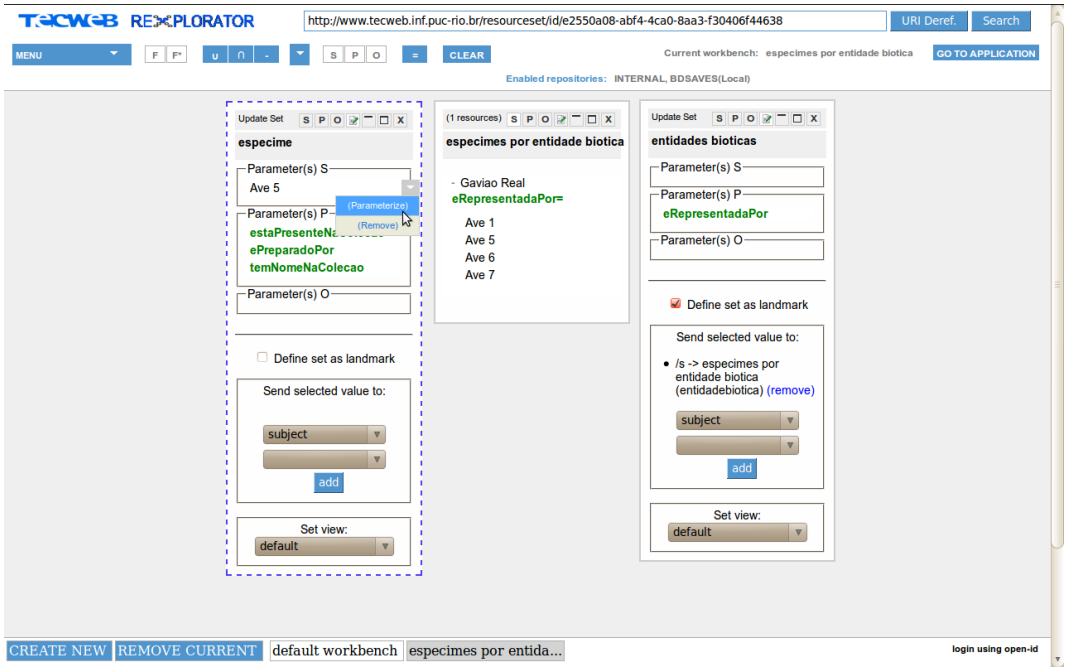


Figura 21 - Parametrizando o conjunto de triplas criado e renomeado

A última tarefa para finalizar a construção do caso de uso é adicionar o comportamento responsável por avaliar o novo conjunto utilizando o espécime selecionado pelo usuário final da aplicação e exibir seus resultados. Para tanto, no modo de edição do conjunto que exibe a lista dos espécimes, deve-se selecionar a posição das triplas na qual o comportamento está sendo adicionado, no caso a posição do objeto. E no segundo *combo-box*, o novo conjunto criado e seu parâmetro para o envio do recurso selecionado pelo usuário final (Figura 22). Assim, está finalizada a implementação do primeiro caso de uso.

Figura 22 - Conjuntos de triplas resultantes que implementam o primeiro caso de uso

Um conceito importante com relação ao compartilhamento fechado é o de consultas subjacentes. O Rexplorator permite o envio de valores para conjuntos de triplas parametrizados de outros *workbenches*, ou seja, que pertencem a outro caso de uso. Esta importante funcionalidade possibilita, no escopo da aplicação aqui gerada, que os conjuntos de visualização dos detalhes de um espécime, por exemplo, sejam acionados sempre que o usuário clicar em uma instância de um espécime, independente dos casos de uso nos quais eles tenham sido criados. Para isso, basta associar o comportamento desejado, de forma análoga ao envio de valores para conjuntos de triplas do mesmo *workbench*. Ainda, para o caso de uso "espécimes por taxon" é possível que quase toda a consulta "espécimes por entidade biótica" seja reutilizada, diminuindo substancialmente o esforço de implementação (Figura 23).

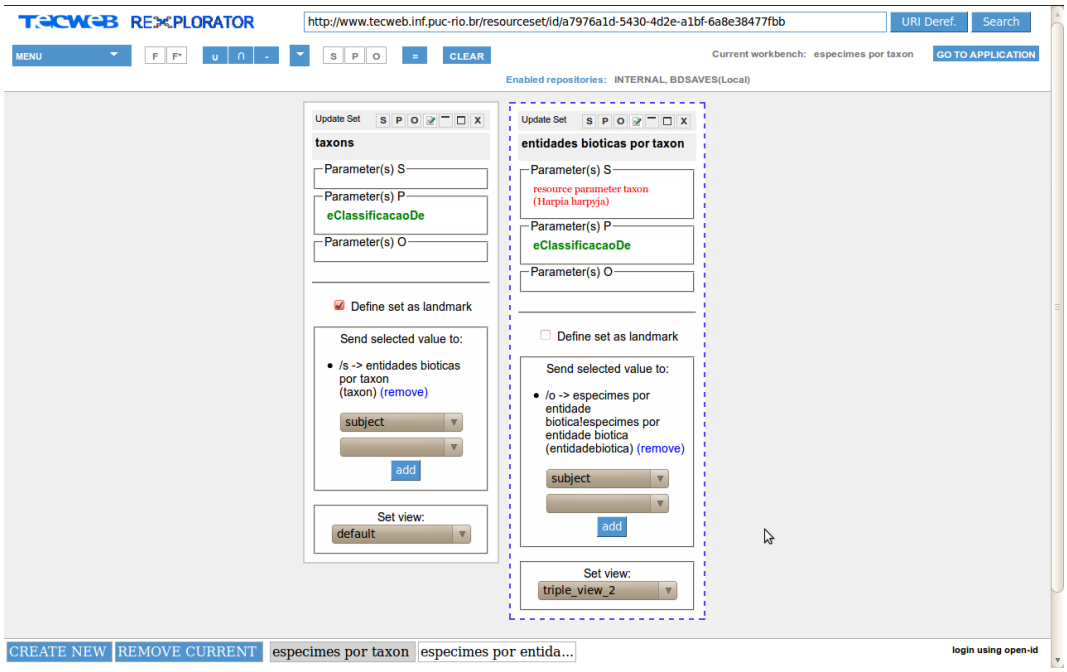


Figura 23 - Conjuntos de triplas resultantes que implementam o segundo caso de uso com o conceito de consulta subjacente

Cada caso de uso da aplicação Web de consulta à base de dados semântica de aves deve ser criado em um *workbench* próprio, nomeado com o nome do caso de uso a qual se refere. O segundo, terceiro, quarto e quinto casos de uso da aplicação Web da secção 4.4.1 foram montados de forma similar ao primeiro, e as suas demonstrações serão omitidas.

O sétimo caso de uso provê para o usuário final uma busca por palavra-chave em um conjunto de nomes científicos dos espécimes presentes na Coleção de Aves do INPA. Para realizar essa busca, é necessário um operador customizado

pode ser utilizado no caso de uso. O código do operador em questão é exibido no quadro abaixo.

```
param_a.select { |triple|
   triple[2].to_s.strip.downcase == param_b[0].to_s.strip.downcase }
```

Quadro 8 - Código do Operador de filtro por palavra-chave

O código é escrito na linguagem ruby. No vetor **param_a** existem as triplas do primeiro parâmetro da operação, enquanto que no vetor **param_b** está o segundo parâmetro, que deve ser uma entrada de texto contendo a palavra-chave a ser buscada. O método **select**, aplicado ao vetor **param_a** possui o comportamento de iterar os elementos do vetor e retornar um novo vetor que possui somente os elementos em que a passagem da iteração retornou o valor primitivo booleano verdadeiro. A variável **triple** ganha o valor da tripla corrente no passo da iteração, representada por um vetor com três posições (sujeito, predicado e objeto). A expressão avaliada em cada passo compara a segunda posição da variável **triple** (objeto da tripla) com o único valor de **param_b**, que é a palavra-chave do transdutor de texto. Um tratamento é dado tanto para o objeto da tripla quanto para a palavra-chave entrada pelo usuário, com o objetivo de aumentar o número de triplas encontradas. A função **to_s** transforma os valores em *string*, a função **strip** retira qualquer espaçamento extra no entorno dos valores e a função **downcase** retira a caixa alta de ambos os valores. Apesar desse operador ter sido construído por Azevedo em [Azevedo, 2011] para ser utilizado em um caso de uso especifico de seu trabalho, ele é genérico o suficiente para filtrar qualquer conjunto de triplas que possua como objeto um literal do tipo *string*. Portanto, ele pôde ser reaproveitado no sexto caso de uso da secção 4.4.1..

Para continuar a construção do caso de uso, é necessário um conjunto com todas as triplas da base que possuem a propriedade temNomeCientifico. Esse conjunto de triplas pode ser criado através de uma operação SPO fixando somente a propriedade temNomeCientifico (Figura 24).

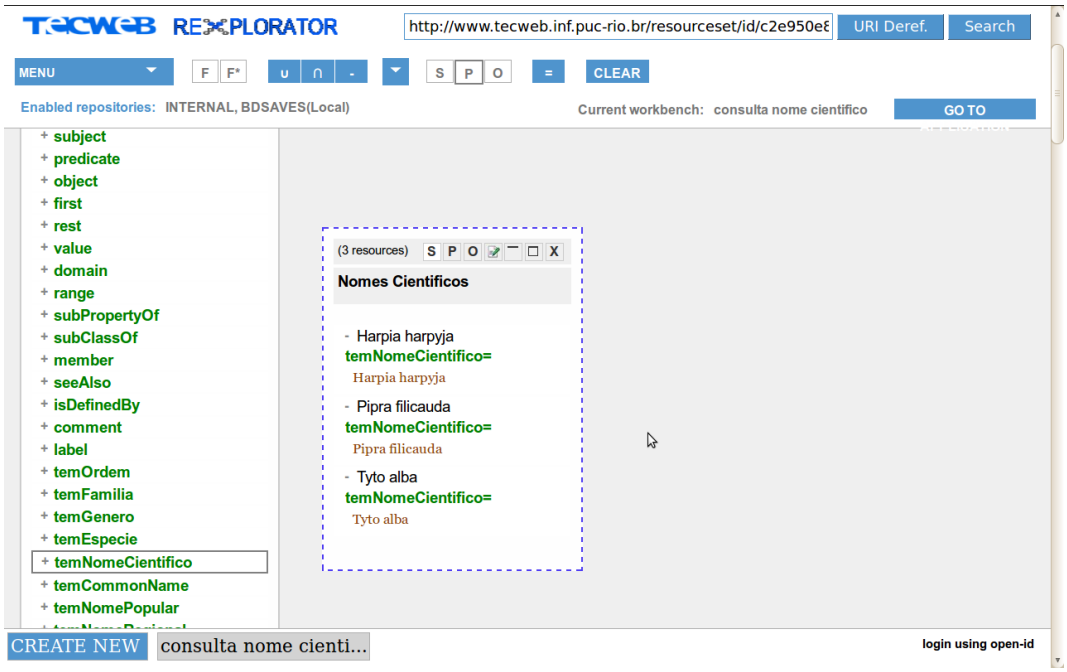


Figura 24 - Criação de um conjunto que possui os táxons na posição de sujeito e os nomes científicos na posição de objeto

Com esse primeiro conjunto criado, é possível utilizar o novo operador para selecionar somente as triplas que possuem a palavra-chave desejada. Para isso é necessário um transdutor para entrada de texto, que pode ser criado através de uma opção no menu principal. O novo transdutor surge no *workbench* e seu nome deve ser redefinido para a semântica desejada. Nesse caso, "Consulta Nome Científico" pode ser utilizado como nome (Figura 25).

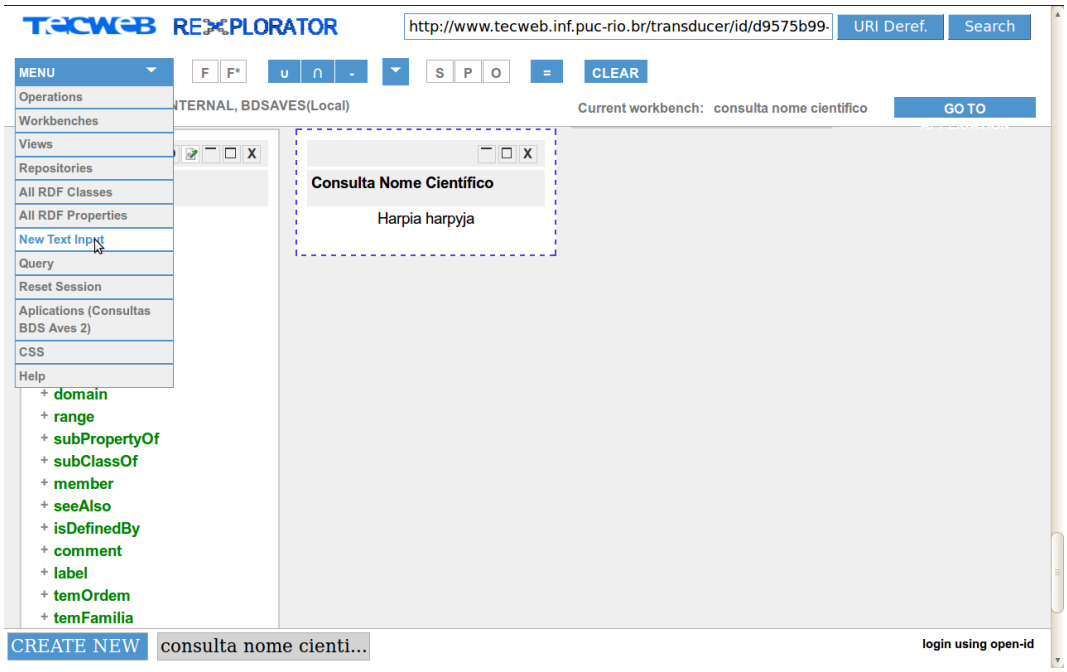


Figura 25 - Opção do menu principal para a criação de uma entrada de texto e transdutor renomeado com a semântica adequada

Com o transdutor criado, é possível construir um segundo conjunto de triplas, que utiliza o primeiro conjunto como primeiro parâmetro e o transdutor de texto como segundo parâmetro (Figura 26). O novo conjunto gerado utiliza o novo operador para filtrar as triplas.

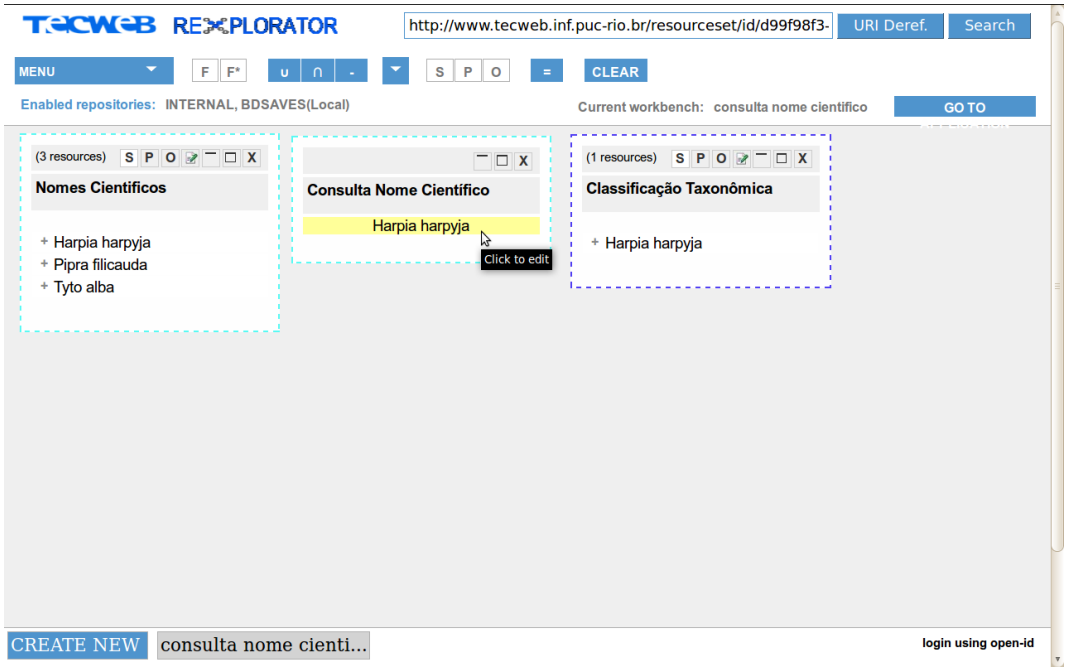


Figura 26 - Conjunto “Classificação Taxonômica” gerado a partir do conjunto “Nomes Científicos” e do transdutor de texto “Consulta Nome Científico”

Uma vez que todos os conjuntos necessários foram gerados, é preciso definir o comportamento da aplicação no caso de uso. O comportamento desejado é que no momento em que o caso de uso for acessado, o sistema exiba o campo de texto para que o usuário faça a consulta pelo nome científico desejado. Depois que o usuário entrar com a palavra-chave, a categoria taxonômica deve ser buscada. O Rexplorator, utilizando os valores fornecidos, calcula o resultado e o exibe na tela. A única tarefa necessária para adicionar o comportamento desejado é, no modo de edição do último conjunto criado (o filtrado pela palavra-chave), selecionar o *checkbox* “*Define set as landmark*” (Figura 27). Uma vez que a opção esteja selecionada, quando o usuário final acessar o caso de uso e fizer sua busca, o Rexplorator vai tentar calcular o valor do conjunto e exibir o resultado.

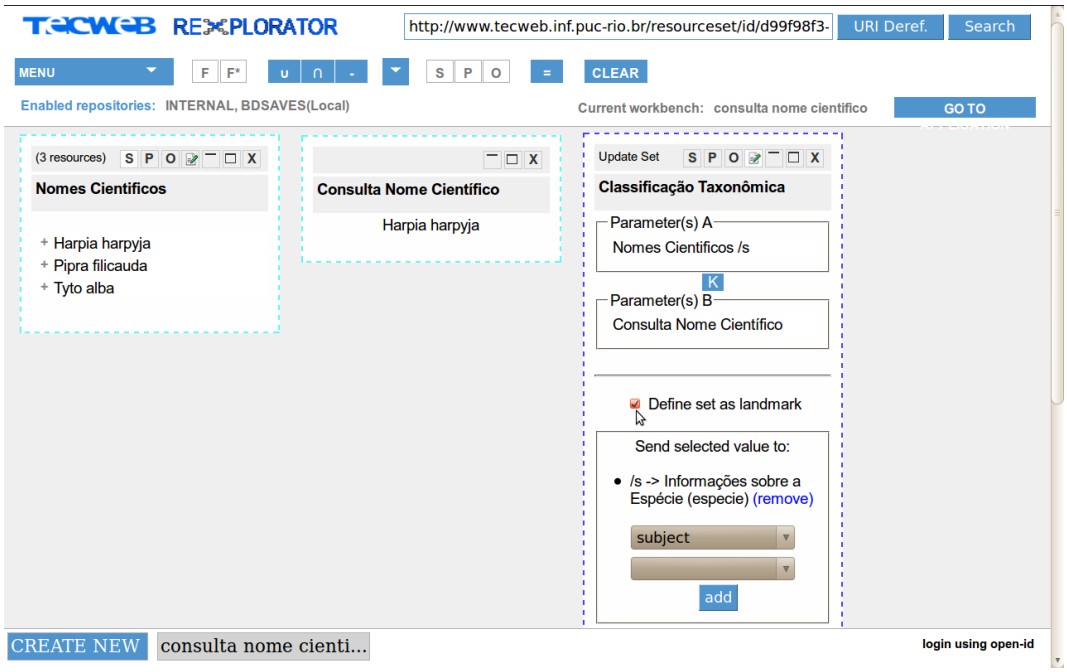


Figura 27 - Conjunto de triplas gerado pela palavra-chave

Seguindo no desenvolvimento do caso de uso, será necessário fornecer informações sobre o nome científico pesquisado. Para isso, é preciso ser criado um conjunto com as propriedades do táxon desejado. Esta tarefa pode ser realizada fixando-se um táxon como sujeito e realizando-se uma consulta SPO. O conjunto gerado possui todas as triplas da base com as propriedades relacionadas ao táxon (Figura 28).

Neste momento, deve-se adicionar o comportamento responsável por avaliar o novo conjunto utilizando o táxon selecionado pelo usuário final e exibir seus resultados. Então, no modo de edição do conjunto que exibe a classificação taxonômica, deve-se selecionar a posição das triplas na qual o comportamento está sendo adicionado, no caso, a posição do sujeito.

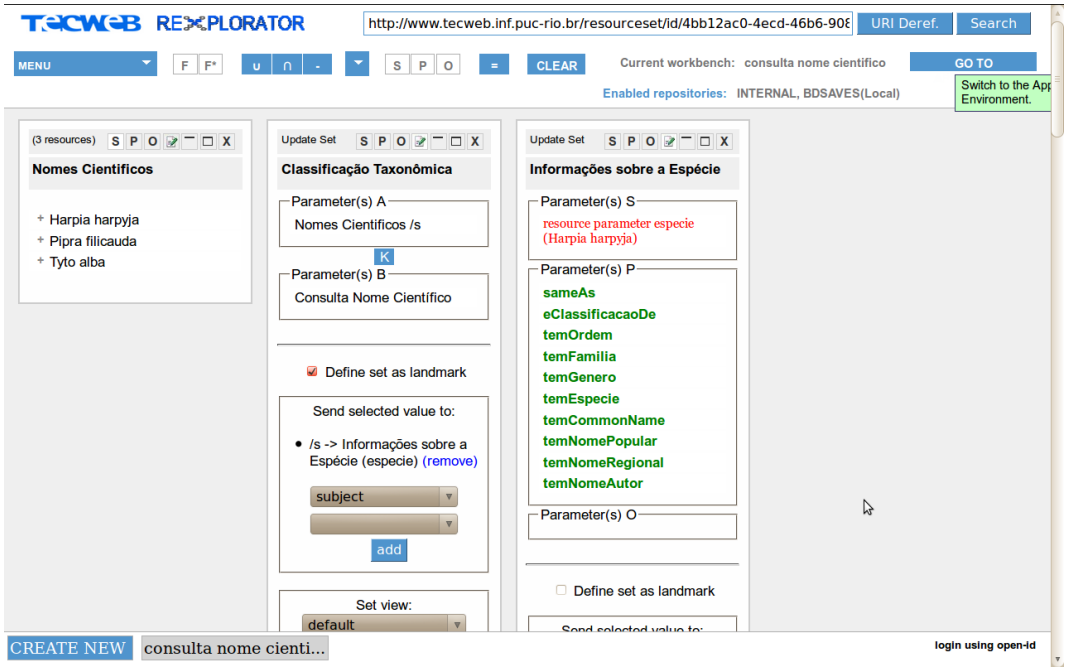


Figura 28 - Conjunto com informações sobre o nome científico pesquisado

O sétimo caso de uso tem um diferencial em relação aos outros. Uma vez que o usuário entrou com um nome científico a ser consultado na base de dados semântica de aves, devem ser apresentadas, além das informações presentes na Coleção de Aves do INPA sobre aquela espécie, informações presentes na LOD.

Para se construir consultas com dados disponibilizados na LOD, precisam ser feitas ligações com recursos em outras bases, conforme o descrito na secção 4.3.6.. O Rexplorator possibilita consultas a SPARQL *Endpoints* no processo de exploração de esquemas RDF. Desta forma, foram adicionados repositórios com informações relevantes para o domínio de pesquisa de aves (Figura 29).

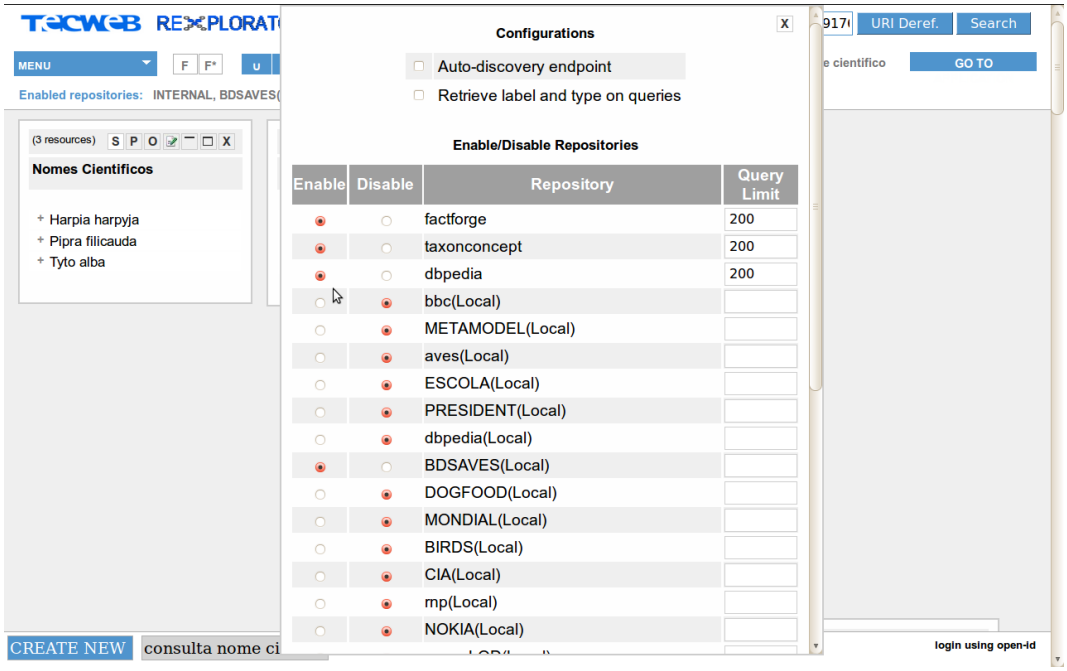


Figura 29 - Inclusão de repositórios da remotos para a geração de consultas

A adição de repositórios permitiu o enriquecimento das consultas criadas para o domínio de aves, uma vez que a propriedade rdfs:sameAs estabeleceu links de identificação para pelo menos três recursos externos à base de dados.

Dando continuidade à implementação do caso de uso, foram gerados conjuntos com informações da DBpedia, da BBC e do *TaxonConcept* para uma dada espécie. Esses conjuntos foram parametrizados, na posição do sujeito, e foram adicionados comportamentos responsáveis por avaliar os novos conjuntos utilizando a espécie selecionada pelo usuário final através da propriedade sameAs no conjunto "Informações sobre espécie". Então, no modo de edição deste conjunto, deve-se selecionar as posições das triplas nas quais os comportamentos estão sendo adicionados, no caso, as posições de objeto (Figura 30).

Vale ressaltar que para cada conjunto com informações externas foi realizada a seleção do que seria exibido para o usuário final objetivando-se a organização da apresentação dos dados e o aproveitamento apenas de informações relevantes.

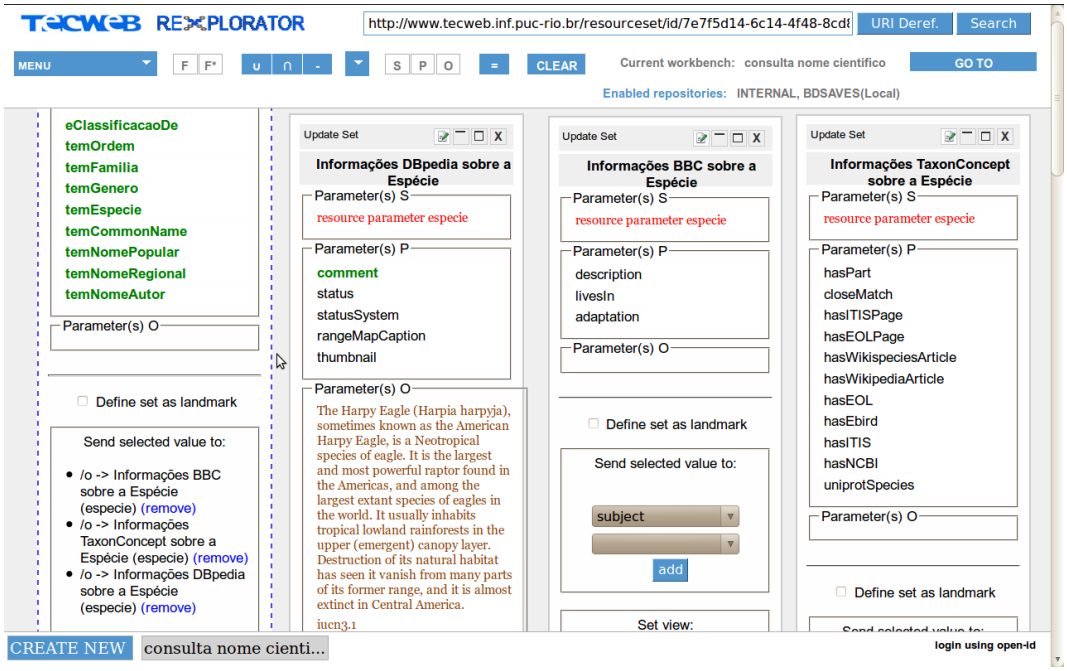


Figura 30 - Adição de informações da LOD através de links de identidade para conjuntos de dados externos

A subsecção a seguir apresenta as telas da aplicação Web desenvolvida neste trabalho onde se pode observar melhor as informações compartilhadas por outras bases.

4.4.3. Telas da aplicação implementada para consultas à base de dados

A seguir, são apresentadas as aplicações de consulta à base de dados semântica de aves implementadas com o Rexplorator e baseadas nos casos de uso descritos anteriormente. Os recursos para a construção de consultas e aplicações do Rexplorator são bem apresentados por Azevedo em [Azevedo, 2010].

Inicialmente, foi utilizado o compartilhamento aberto de consultas definido por Azevedo para a verificação das necessidades dos usuários com relação às consultas feitas à Coleção de Aves do INPA. Em seguida, foi utilizado o compartilhamento fechado para a criação de uma aplicação para a mesma coleção.

Na Figura 31, é apresentada a tela inicial da aplicação com botões representando os seis primeiros casos de uso descritos na secção 4.4.1..

Figura 31 - Tela inicial da aplicação de consulta à Coleção de Aves do INPA

Espécimes por Entidade Biótica

Ao se clicar no botão “especimes por entidade biotica” no menu principal, (Figura 32 - (1)), uma lista com as entidades bióticas disponíveis na base aparece para a escolha pelo usuário (Figura 32 - (2)). No exemplo, é escolhida a entidade biótica “Gaviao Real Femea”.

Figura 32 - (1) Seleção da consulta de espécimes por entidade biótica; (2) Seleção da entidade biótica da lista de entidades disponíveis

Ao ser selecionada a opção de entidade biótica, a aplicação apresenta os espécimes disponíveis na base de dados que representam aquela entidade específica (Figura 33 - (3)). Neste momento, pode-se selecionar um dos espécimes (no exemplo, foi selecionada a “Ave 4”) para se consultar mais informações na base sobre o mesmo; no caso, quem foi o preparador do espécime e em quais coleções o mesmo consta (Figura 33 - (4)).

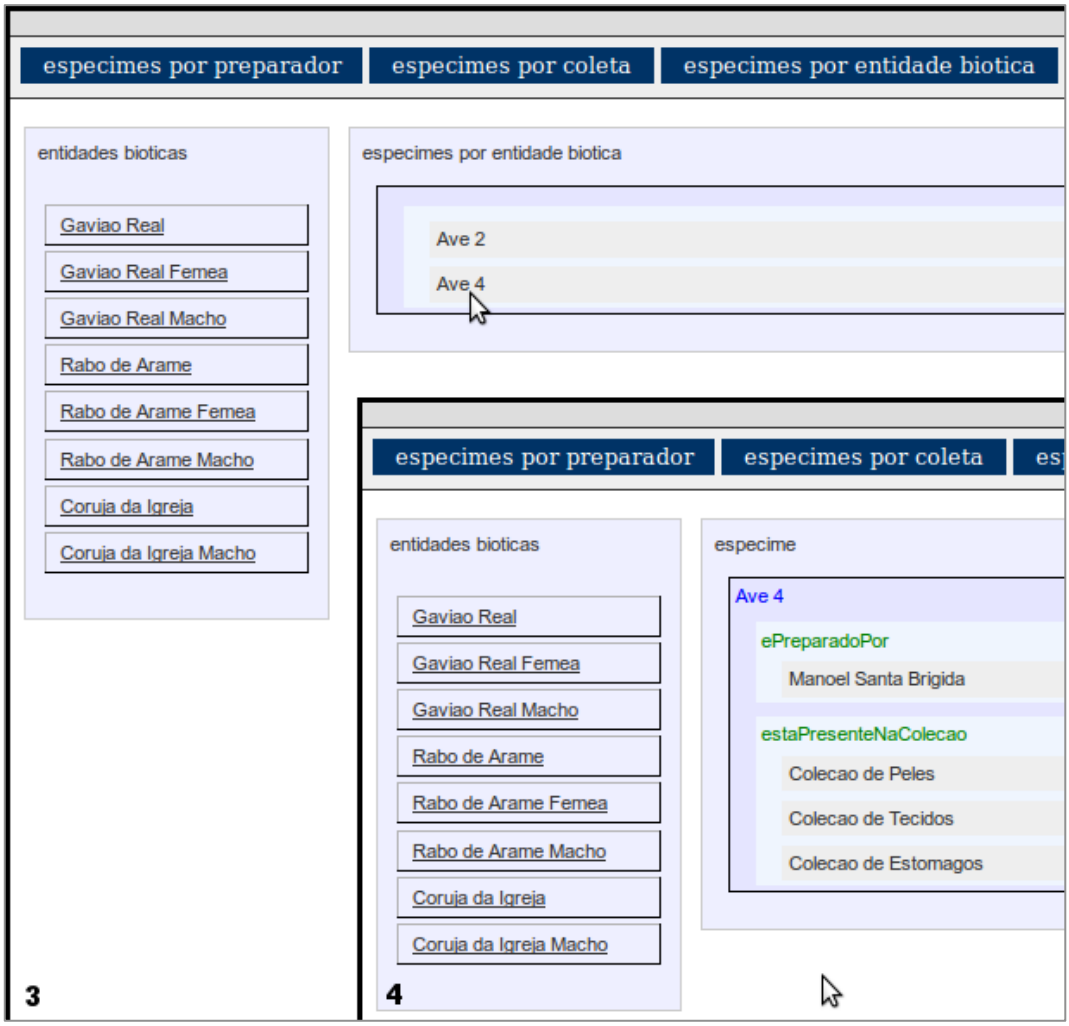


Figura 33 - (3) Espécimes por entidade biótica; (4) Informações adicionais sobre o espécime

Espécimes por Táxon

Ao se clicar no botão “especimes por taxon” no menu principal, (Figura 34 - (1)), uma lista com as categorias taxonômicas disponíveis na base aparece para a escolha pelo usuário (Figura 34 - (2)). No exemplo, é escolhido o táxon *Pipra filicauda*.

Ao ser selecionada a opção de táxon, a aplicação apresenta as entidades bióticas disponíveis na base de dados classificadas por aquele táxon específico (Figura 34 - (3)). Neste momento, pode-se selecionar uma das entidades para se consultar informações sobre os espécimes representados por aquela entidade na base de dados semântica de aves. No exemplo, foi selecionada a entidade “Rabo de Arame Macho”.

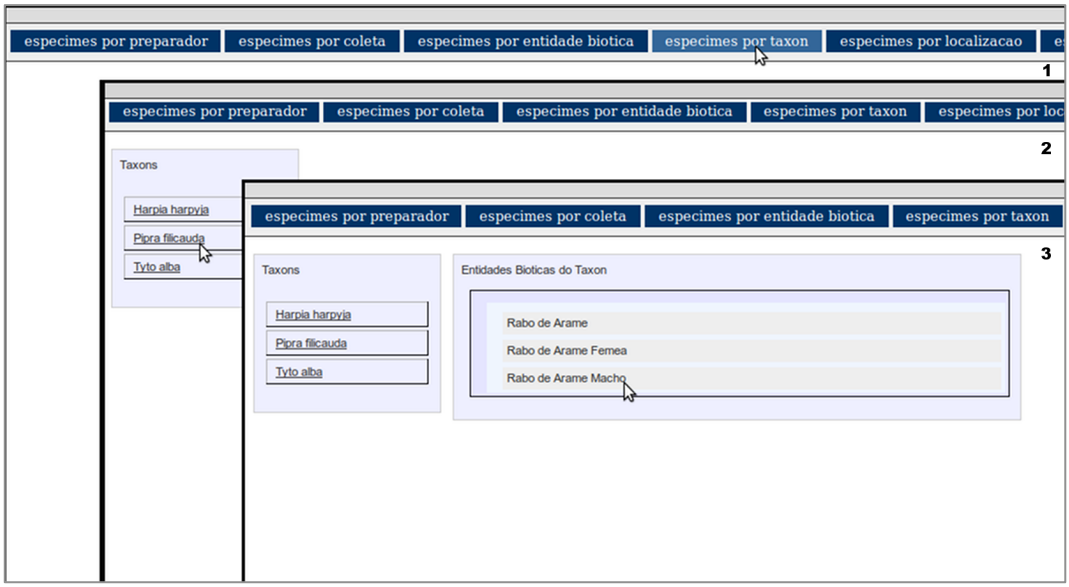


Figura 34 - (1) Seleção da consulta de espécimes por táxon; (2) Seleção do táxon da lista de categorias taxonômicas disponíveis; (3) Seleção da entidade biótica

Ao ser selecionada a opção de entidade e apresentados os espécimes disponíveis que a representam específica (Figura 35 - (4)), pode-se selecionar um dos espécimes (no exemplo, foi selecionada a "Ave 15") para se consultar mais informações na base sobre o mesmo; no caso, quem preparou o espécime, em qual coleção está presente e qual seu nome na coleção de tecidos (Figura 35 - (5)).

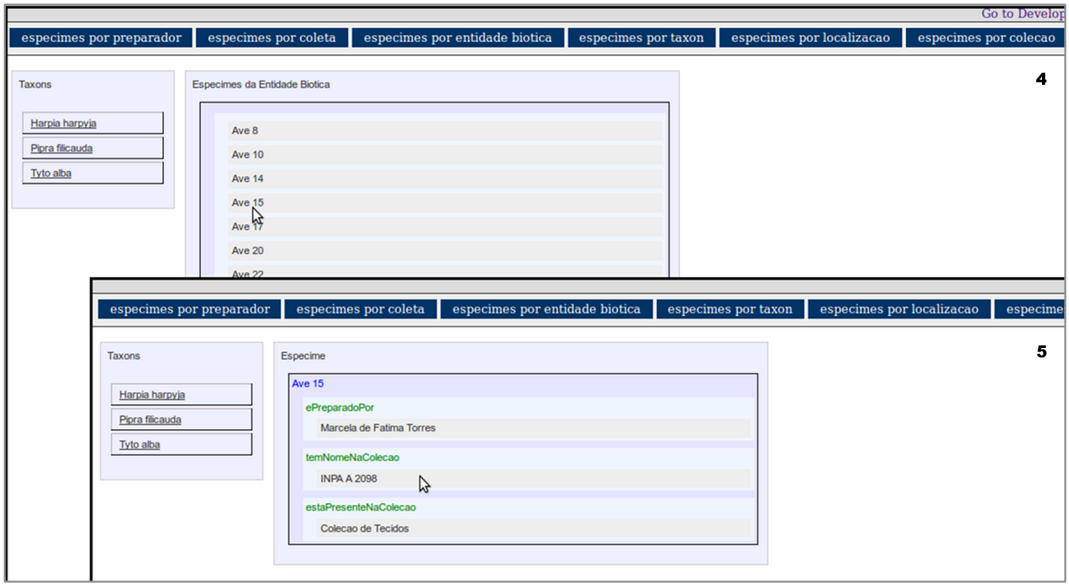


Figura 35 - (4) Espécimes por entidade biótica; (5) Informações adicionais sobre o espécime

Espécimes por Coleção

Ao se clicar no botão "especimes por colecao" no menu principal, (Figura 36 - (1)), uma lista com os tipos de coleção de aves disponíveis na base aparece

para a escolha pelo usuário (Figura 36 - (2)). No exemplo, é escolhida a “Colecao de Peles”.

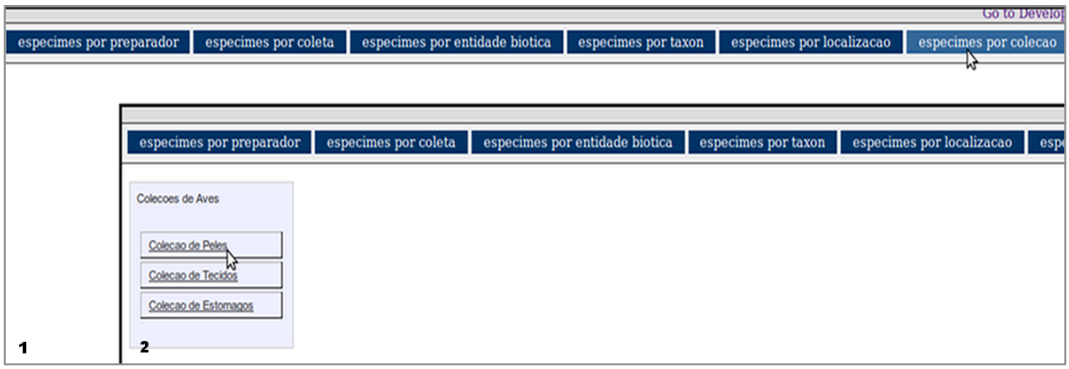


Figura 36 - (1) Seleção da consulta de espécimes por coleção; (2) Seleção da Coleção de Peles da lista de coleções disponíveis

Ao ser selecionada a opção de coleção, a aplicação apresenta os espécimes disponíveis na base de dados presentes naquela coleção específica (Figura 37 - (3)). Neste momento, pode-se selecionar um dos espécimes (no exemplo, foi selecionada a “Ave 7”) para se consultar mais informações na base sobre o mesmo; no caso, quem preparou o espécime e qual o nome do mesmo na coleção de peles (Figura 37 - (4)).

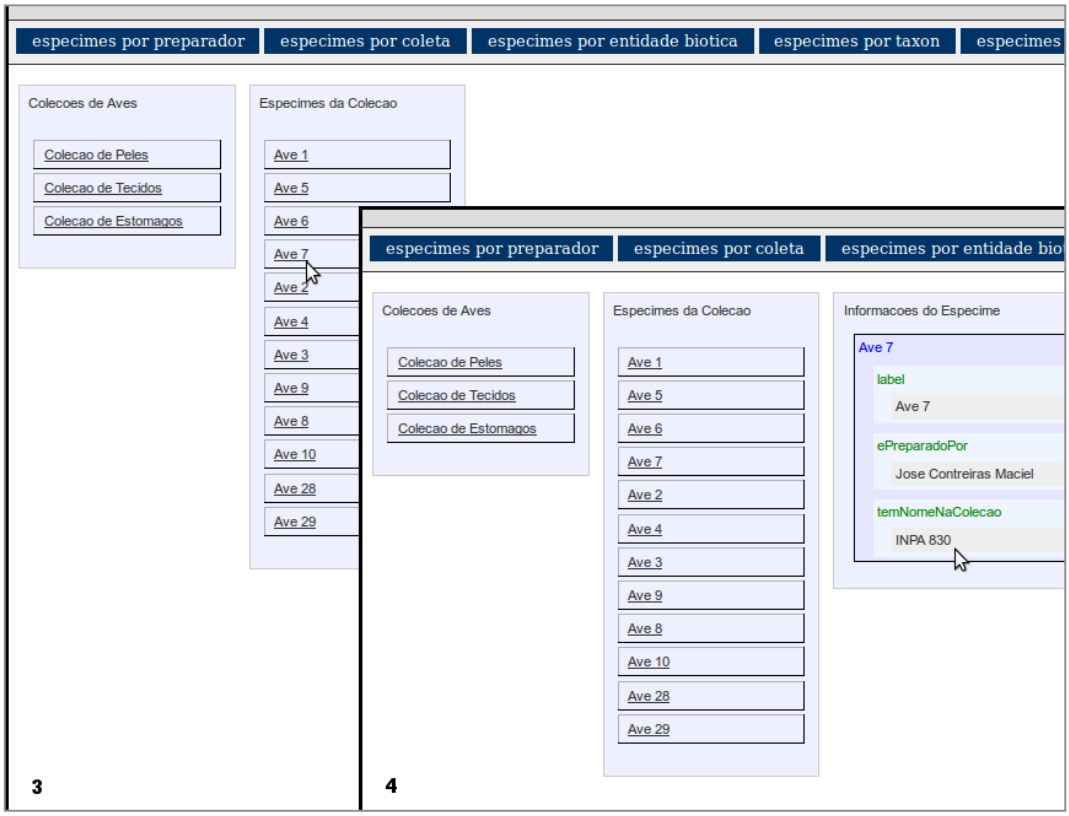


Figura 37 - (3) Espécimes por coleção; (4) Informações adicionais sobre o espécime

Espécimes por Preparador

Ao se clicar no botão “especimes por preparador” no menu principal, (Figura 38 - (1)), uma lista com os preparadores de espécimes para a coleção disponíveis na base aparece para a escolha pelo usuário (Figura 38 - (2)). No exemplo, é escolhido o preparador “Mario Cohn-Haft”.

Figura 38 - (1) Seleção da consulta de espécimes por preparador; (2) Seleção do preparador da lista de preparadores disponíveis

Ao ser selecionada a opção de preparador, a aplicação apresenta os espécimes disponíveis na base de dados preparados por aquele preparador específico (Figura 39 - (3)). Neste momento, pode-se selecionar um dos espécimes (no exemplo, foi selecionada a “Ave 10”) para se consultar mais informações na base sobre o mesmo; no caso, em quais coleções o espécime está presente e qual o nome que ele recebe em cada coleção (Figura 39 - (4)).

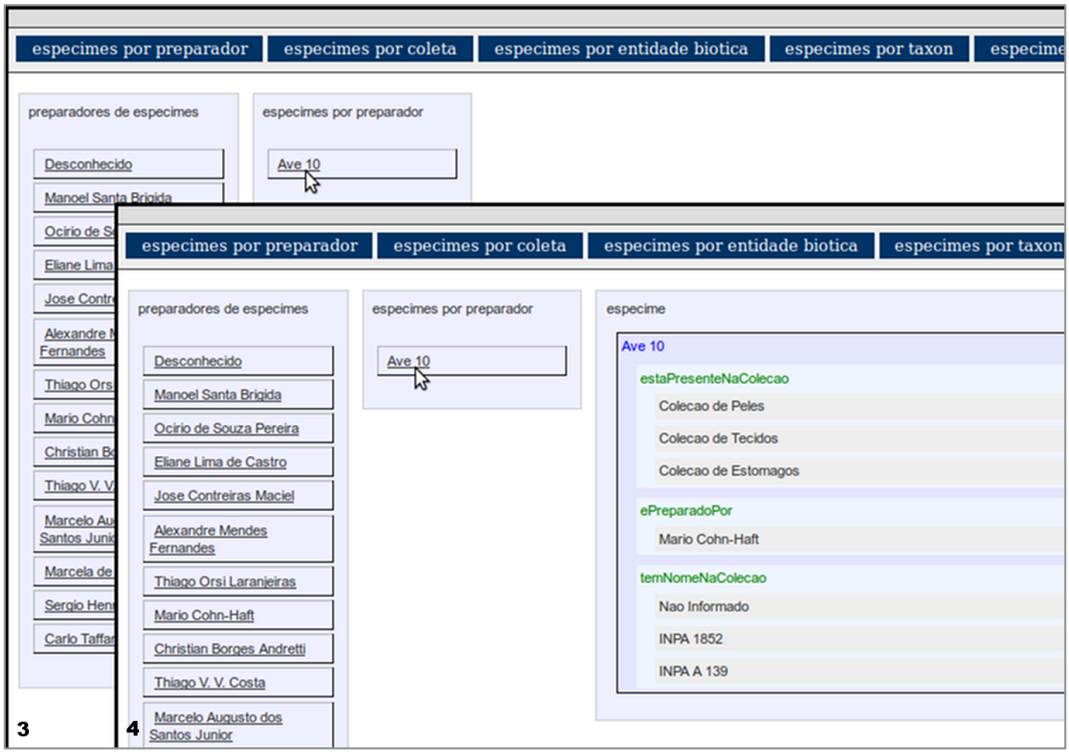


Figura 39 - (3) Espécimes por preparador; (4) Informações adicionais sobre o espécime

Espécimes por Coleta

Ao se clicar no botão “especimes por coleta” no menu principal, (Figura 40 - (1)), uma lista com as coletas disponíveis na base aparece para a escolha pelo usuário (Figura 40 - (2)). No exemplo, é escolhida a “Coleta 14”.

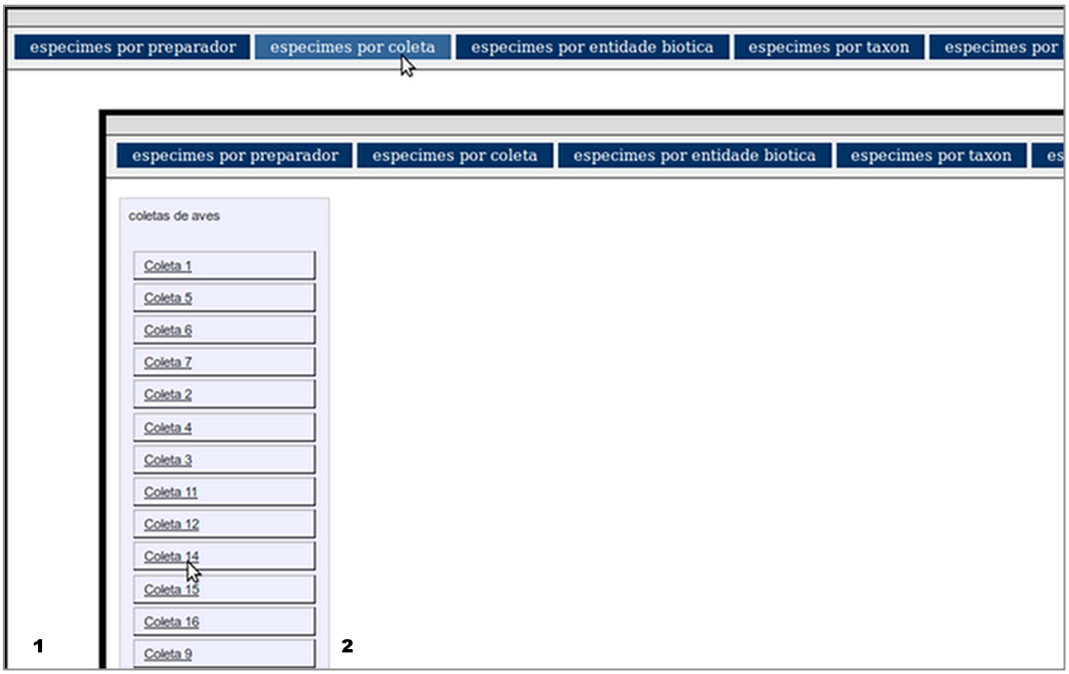


Figura 40 - (1) Seleção da consulta de espécimes por coleta; (2) Seleção de coleta da lista de coletas disponíveis.

Ao ser selecionada uma coleta, a aplicação apresenta os espécimes disponíveis na base de dados que coletados durante a mesma (Figura 41 - (3)). Neste momento, pode-se selecionar um dos espécimes (no exemplo, foi selecionada a "Ave 21") para se consultar mais informações na base sobre o mesmo; no caso, quem foi o preparador e em quais coleções está presente (Figura 41 - (4)).

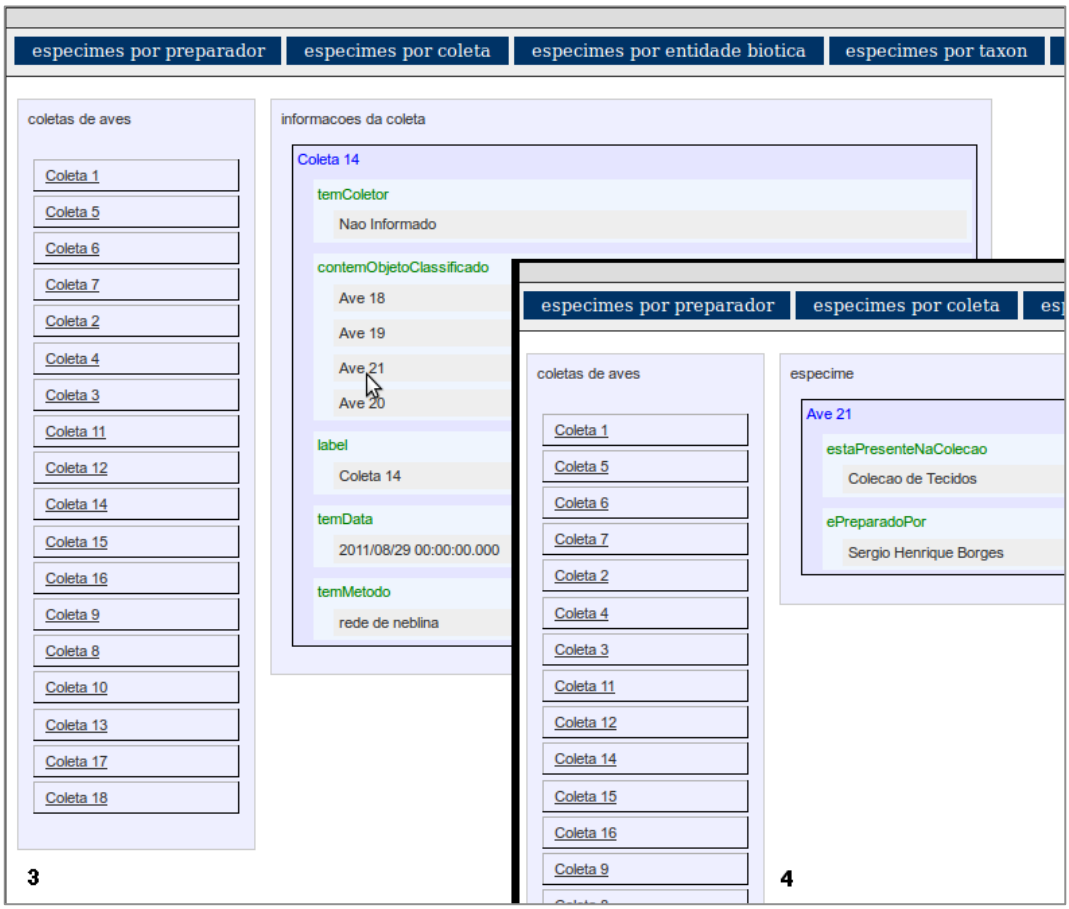


Figura 41 - (3) Espécimes por coleta; (4) Informações adicionais sobre o espécime

Espécimes por Localização

Ao se clicar no botão "especimes por localizacao" no menu principal, (Figura 42 - (1)), uma lista com os locais de coleta disponíveis na base aparece para a escolha pelo usuário (Figura 42 - (2)). No exemplo, é escolhida a reserva florestal do INPA denominada "ZF2". Após a seleção, sistema apresenta as informações existentes na base referentes à localização selecionada (Figura 42 - (3)). No sentido de buscar espécimes para aquela localização, uma informação relevante nesta última tela é indicação de quais coletas ocorreram naquele local. No exemplo, é selecionada a "Coleta 7", também na Figura 42 - (3).

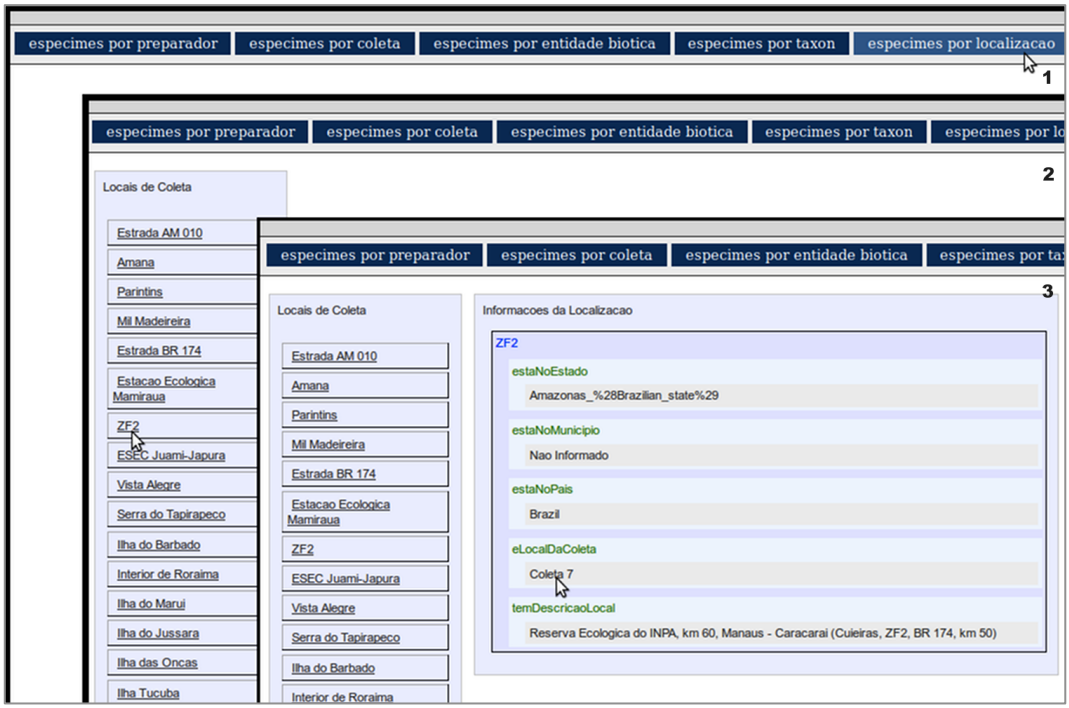


Figura 42 - (1) Seleção da consulta de espécimes por localização; (2) Seleção do local da lista de localizações disponíveis; (3) Seleção de coleta realizada na localização

Ao ser selecionada a coleta de interesse, a aplicação apresenta os espécimes disponíveis na base de dados coletados na mesma (Figura 43 - (4)). Neste momento, pode-se selecionar um dos espécimes (no exemplo, foi selecionada a "Ave 7") para se consultar mais informações na base sobre o mesmo (Figura 43 - (5)); no caso, a entidade biótica que ele representa na coleção.

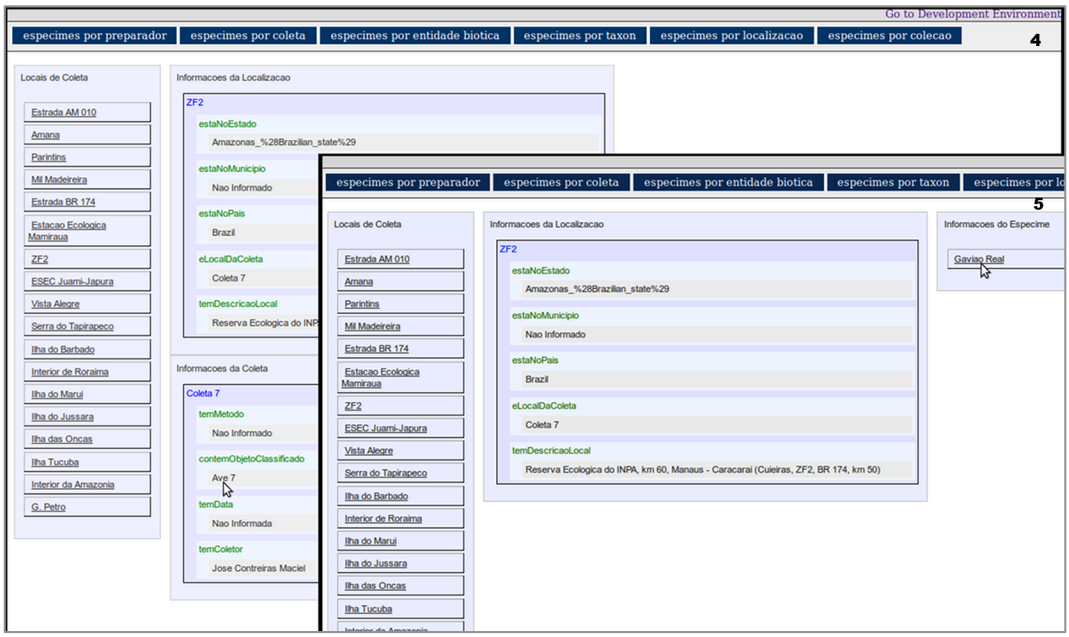


Figura 43 - (4) Espécimes por localização; (5) Informações adicionais sobre o espécime

Consulta Nome Científico

Ao se clicar no botão “consulta nome cientifico” no menu principal, (Figura 44 - (1)), é exibido um campo de texto para que o usuário entre com o nome científico como palavra-chave para consulta na base de dados (Figura 44 - (2)). Depois que o usuário entra com o nome desejado, se houver a espécie procurada na coleção de aves, é exibido como resultado uma tripla com o nome pesquisado (Figura 44 - (3)). No exemplo, é realizada uma busca pelo nome científico *Harpia harpyja*.

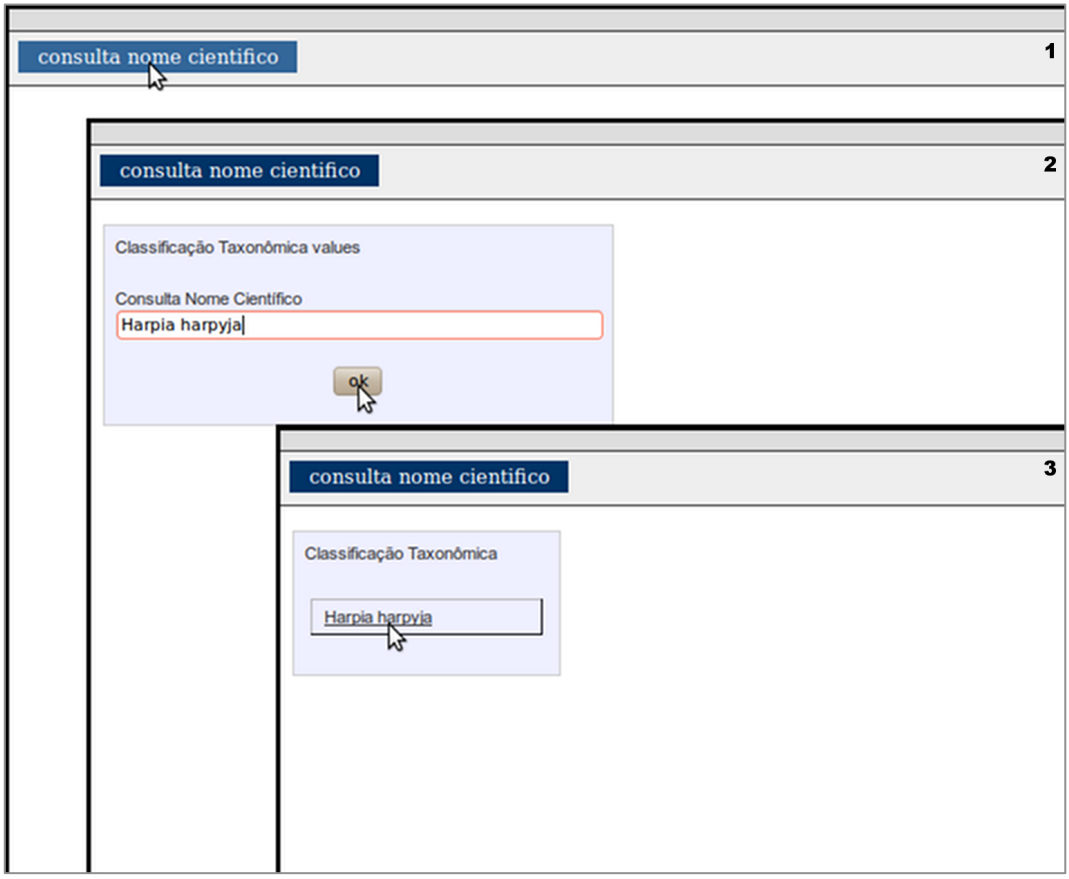


Figura 44 - (1) Seleção da consulta por nome científico, (2) Busca por nome científico e (3) Tripla resultante da consulta por nome científico

Ao ser selecionada a tripla com o nome científico, a aplicação apresenta informações da base de dados semântica local sobre o mesmo (Figura 45 - (4)). Neste momento, pode-se selecionar um dos recursos da LOD ligados a um recurso da base local pela propriedade rdfs:sameAs. Ainda na tela 4 da Figura 45, pode ser observada a seleção da opção “Harpy_Eagle”, como recurso “igual” ao nome científico procurado na consulta inicial. Ao ser selecionada essa opção, as informações existentes na Dbpedia sobre o nome científico *Harpia harpyja* são recuperadas e exibidas para o usuário (Figura 45 - (5)). É importante observar,

que, apesar de serem informações gerais sobre o nome científico procurado, elas complementam as existentes na base local. Ainda, algumas são particularmente interessantes para os ornitólogos, como, por exemplo, saber qual a categoria (*status*) de conservação de uma espécie na classificação dada pelo *International Union for Conservation Nature* - IUCN[71]. No caso do *Harpia harpyja*, o *status* IUCN é, de acordo com o sistema de classificação IUCN versão 3.1 e através da DBpedia, *Near Threatened* (NT), que significa quase ameaçada de extinção.

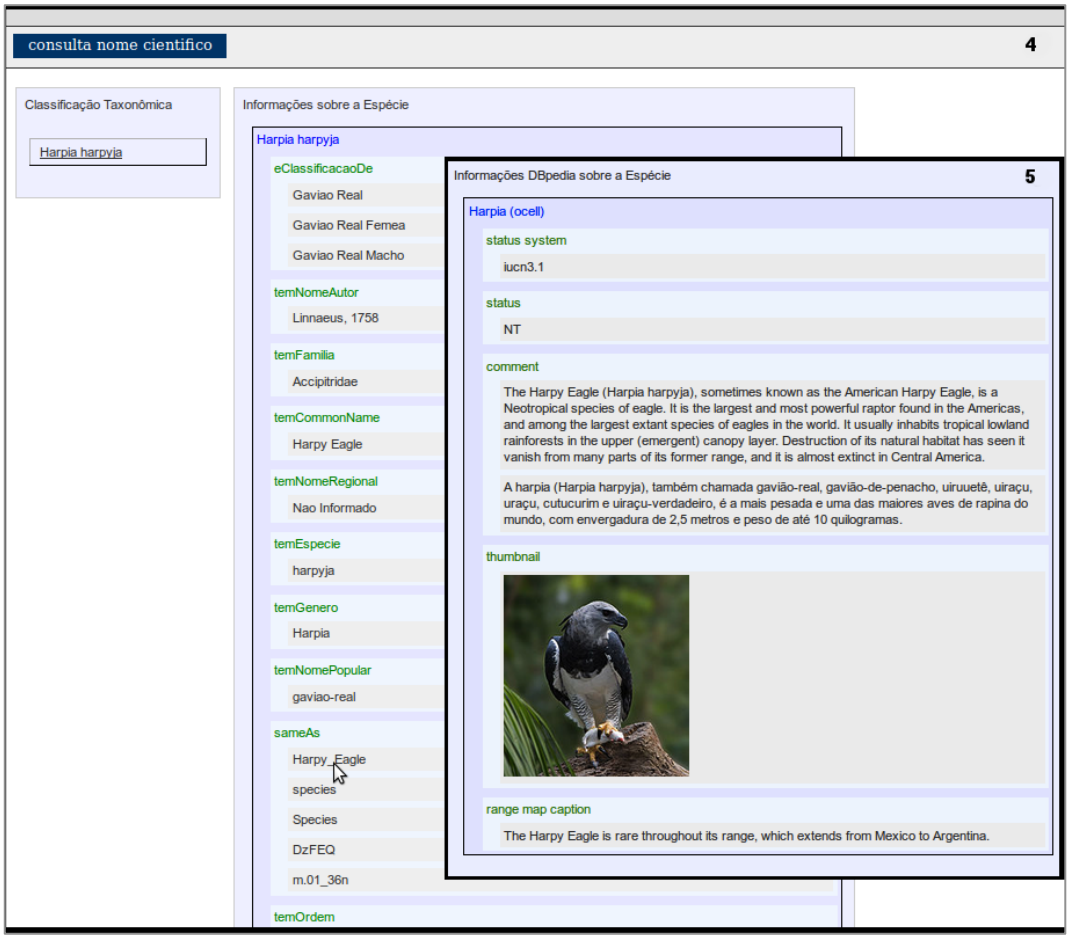


Figura 45 - (4) Espécimes por entidade biótica; (5) Informações sobre o espécime

Analogamente, a Figura 46 - (6) apresenta a lista de informações da base local sobre a espécie *Harpia harpyja* e a seleção da propriedade rdfs:sameAs "*Harpia harpyja* (Linnaeus 1758)", como recurso também "igual" ao nome científico procurado na consulta inicial. Ao ser selecionada essa opção, as informações existentes no projeto TaxonConcept sobre o nome científico *Harpia harpyja* são recuperadas e exibidas para o usuário (Figura 46 - (7)). Neste caso, são referências diretas ao nome científico procuradas pelos pesquisadores em sites

[71] http://www.iucn.org/

especializados sobre ornitologia como a do Ebird[72] (no caso, o código Ebird "hareag1"), que possui informações sobre a abundância de aves e sua distribuição global em várias escalas espaciais e temporais; e a do GenBank[73] (no caso, o código no GenBank "202280"), uma base de dados com a sequência genética das espécies que auxilia os ornitólogos na identificação de espécimes coletados e na constatação de descoberta de novas espécies.

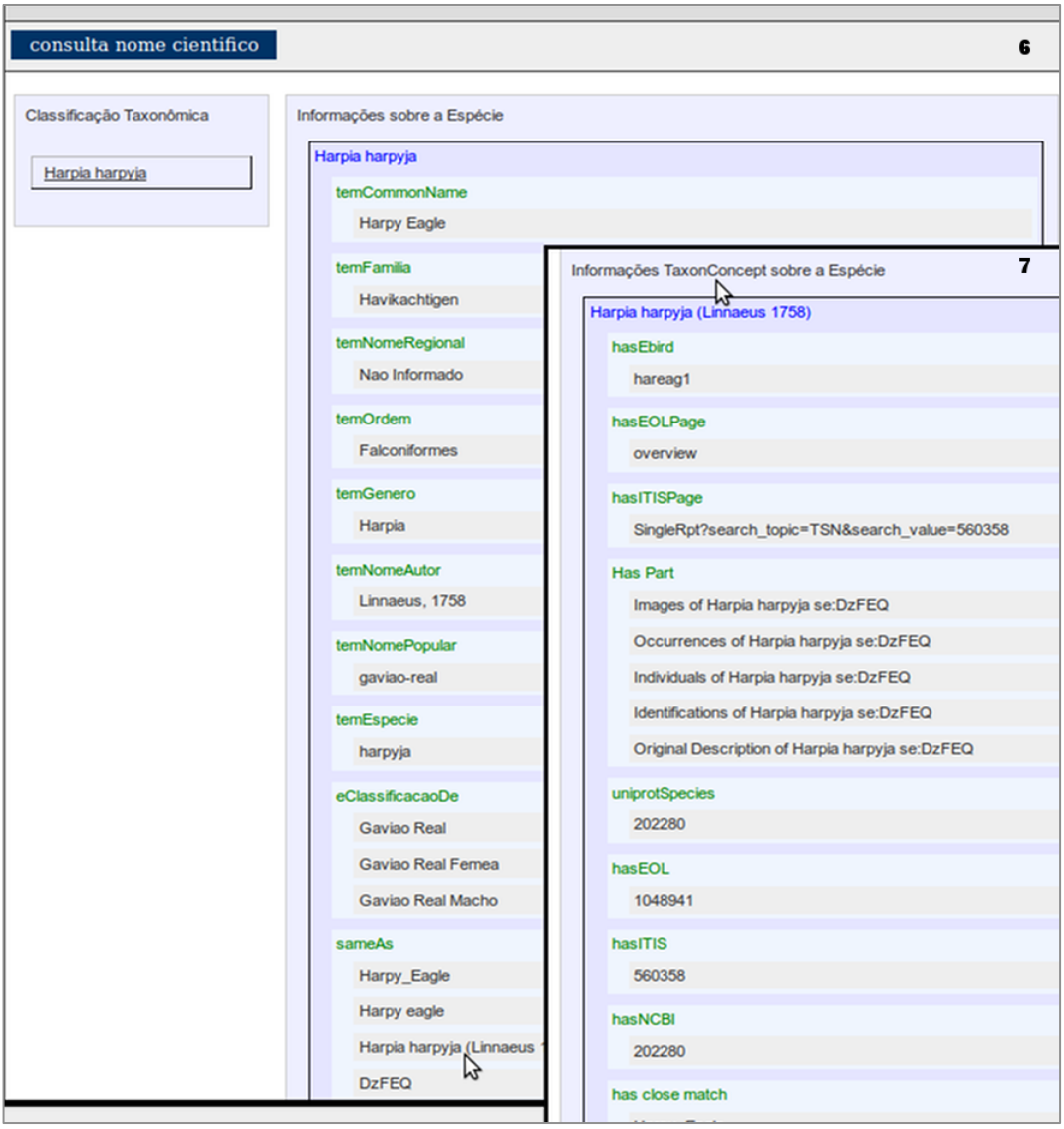


Figura 46 - (6) Lista de informações da base local sobre a espécie e seleção de recurso da base TaxonConcept; (7) Informações TaxonConcept sobre a espécie

[72] http://ebird.org/content/ebird/
[73] http://www.ncbi.nlm.nih.gov/genbank/

4.5.
Impacto nos processos de pesquisa

Quando se tenta praticar a gestão do conhecimento científico, é necessário entender a maneira pela qual o conhecimento é obtido, quem o possui, como ele está formatado e que barreiras, físicas e culturais, devem ser transpostas para codificá-lo e disseminá-lo.

Neste trabalho pôde-se perceber que para avaliar o impacto da aplicação de métodos e ferramentas modernas de tecnologia da informação, leia-se as tecnologias da Web Semântica, será necessário a definição de um protocolo com métricas reais de avaliação das várias etapas da pesquisa.

O grupo de pesquisa em ornitologia do INPA, liderados pelo Dr. Mário Cohn-Haft, desenvolve uma metodologia particular de coleta e aproveitamento do material biótico coletado que deve ser bem mapeada para a identificação de que pontos outros podem ser auxiliados pelas novas tecnologias de TI.

Ainda assim, a realização deste trabalho permitiu evidenciar algumas vantagens da aplicação das ferramentas baseadas na Web Semântica na coleção biológica de aves:

1. Um processo de curagem de dados digitais permite aos pesquisadores que elaboram as planilhas eletrônicas identificarem erros primários em seus conjuntos de dados, como a grafia incorreta de nomes científicos ou mesmo erros de medidas. Porém, permite perceberem os benefícios que uma ferramenta de limpeza, correção e normalização de dados, como o Google Refine, pode oferecer em um processo de validação e certificação dos dados científicos visando um futuro serviço de proveniência e garantindo a qualidade dos resultados dos processos de análise e síntese desses dados.
2. A estruturação dos dados digitalizados para o consumo pela WS possibilitou aos pesquisadores observarem a praticidade e agilidade na recuperação da informação desejada e adquirida partindo-se dos mesmos dados que utilizam rotineiramente.
3. Passaram a considerar associações com outros domínios que não tinham pensado anteriormente. Um exemplo disso, foi considerar a localização de uma espécie com base na análise genômica do

conteúdo estomacal dos espécimes de coleção, uma vez que a coleção de estômagos tem o potencial de revelar a dieta vegetal de uma determinada espécie e, consequentemente, sua localização considerando a ocorrência de plantas em determinado habitat. Outro exemplo, foi considerar que uma espécie pode ser localizada com base no tipo de árvore em que costuma fazer ninho, pois, uma vez localizada aquela espécie de árvore, seria possível a verificação de sua ocorrência. Esses tipos de associação podem influenciar tomadas de decisão como, por exemplo, o planejamento de uma expedição para localizar determinada espécie ameaçada de extinção.

4. Mudou-se a forma de enxergar as coleções na medida em que uma coleção de aves, por exemplo, não está restrita ao domínio de aves, mas o extrapola através da verificação de que existem associações com outras coleções biológicas e também com outros domínios disponíveis na Web. Este novo olhar permitirá o aumento na capacidade de geração de informação e de conhecimento cuja mensuração pode até ser possível, mas foge ao escopo desse trabalho.
5. O grupo de pesquisa entendeu a importância de se agregar semântica aos dados ao experimentarem a recuperação da informação desejada com a possibilidade de ligação com outras bases da LOD, tornando seu processo de pesquisa na Web automático, rápido e uma aquisição de conhecimento mais completa. Ainda, o grupo descobriu novas oportunidades de ligação na LOD como, por exemplo, a recuperação do código de referência no GenBank, informação adquirida via ligação com o Projeto *TaxonConcept* e que é utilizada para a verificação do sequenciamento genômico de uma espécie, a correção de erros na base do GenBank ou mesmo a verificação de descoberta de uma espécie nova.
6. A nova visão de que o conhecimento sobre a Coleção de Aves da Amazônia do INPA pode ser expandido, persistido e compartilhado, mesmo após o fim do ciclo de pesquisa com o atual grupo de ornitólogos, que agora é facilitada com a utilização de tecnologias da WS.

5
Considerações Finais

A transdisciplinaridade da ciência evidencia a computação como recurso indispensável para a gestão e monitoramento dos recursos naturais. A recente formação de uma comunidade de informática dedicada a solucionar problemas relativos à biodiversidade envida a descrição (sintático e semântica) e acesso a qualquer informação complementar que possa estar associada a registros de coleta de um espécime ou sua ocorrência. Tal informação amplia o escopo de dados potencialmente relevantes para incluir um conjunto de medidas observadas sobre os aspectos bióticos e abióticos do ambiente. Por exemplo, ao analisar padrões na abundância global e distribuição espacial de certos táxons, informações sobre a situação da co-precipitação, umidade, tipo de solo, uso da terra, nível do rio, etc., poderiam ser parâmetros importantes a serem considerados. Assim, a necessidade de integração da referida comunidade finalmente converge com as de outros domínios, que dependem de dados multifacetados para uma maior compreensão. Esta amplitude necessária passa pela adoção de novas tecnologias que assumem um universo semântico a ser tratado e explorado isentando o usuário final das complexidades inerentes.

Além disso, o grande volume de conhecimento gerado por centros de pesquisa científica demanda a gestão desse conhecimento produzido pelos pesquisadores de modo a proporcionar ambientes e ferramentas colaborativos, possibilitando a geração de inovações e novos conhecimentos visando atender às demandas da sociedade.

Este trabalho proporcionou a aquisição de conhecimentos acerca de processos de pesquisas desenvolvidos no INPA, especificamente o Programa de Coleções Biológicas, e em instituições similares na Amazônia, por exemplo, Museu Paraense Emílio Goeldi - MPEG e Universidade Federal do Amazonas - UFAM.

Tradicionalmente, o ciclo de pesquisa compreende as fases de coleta, curagem, análise e visualização de dados. As coleções biológicas representam,

portanto, um acervo de conhecimento científico a ser utilizado no ciclo de pesquisa e que demanda gestão física, de dados e metadados.

5.1. Resultados alcançados

Os principais resultados alcançados são destacados a seguir:

- O estudo e constatação de um paralelo entre a evolução da Web e a evolução das Coleções Biológicas apresentado Capítulo 2.
- A descrição de um processo de criação de bases de dados semânticas, manipulação destes dados e desenvolvimento de aplicações Web Semânticas com base nas tecnologias mais recentes relatado no Capítulo 3 desta dissertação.
- A realização de um Estudo de Caso com dados reais sobre aves da Amazônia apresentado no Capítulo 4. Esta aplicação da metodologia descrita no Capítulo 3, comprovou a validade do processo descrito, bem como possibilita sua reutilização para os demais domínios de pesquisa com bases de dados científicas do INPA.
- A comprovação dos benefícios causados pela aplicação de tecnologias da Web Semântica nas coleções biológicas do INPA através da consulta a pesquisadores especialistas em ornitologia.

5.2. Contribuições do trabalho

Considerando-se a complexidade do universo biológico rico em semântica e os avanços da Web Semântica, que demonstram recursos para mitigar questões de interoperabilidade, integração e gestão dos dados, o trabalho permitiu o estudo do estado da arte das tecnologias desenvolvidas para: construção de bases de dados semânticas (Google Refine, Kettle, D2R Server, Silk, Sesame); exploração de dados em RDF e construção de aplicação Web para a WS (Rexplorator).

O trabalho contribuiu no processo de extração da semântica dos dados digitalizados da Coleção de Aves do INPA. Os dados sobre aves encontram-se digitalizados em planilhas com informação específica do domínio de ornitologia e seguindo a orientação do Manual de Taxidermia da Coleção de Aves do INPA.

Após entrevistas com os pesquisadores especialistas em ornitologia e consulta ao referido manual, informações sobre taxonomia, localização e preparo de acervo científico puderam ser compreendidas e utilizadas no processo de construção da base de dados semântica sobre aves da coleção.

A realização do trabalho proporcionou a estruturação dos dados científicos da Coleção de Aves do INPA, com a realização de anotação semântica dos mesmos de modo a permitir o processamento por máquinas e, consequentemente, agilizando procedimentos de consulta à base de dados sobre esta coleção.

Ainda, o emprego de vocabulários utilizados na LOD na composição da base de dados semântica sobre aves, permite a futura ligação desses dados com bases similares disponíveis na Web Semântica, promovendo a colaboração entre grupos de pesquisa do domínio da ornitologia, bem como a colaboração com grupos de pesquisa de domínios diferentes e que considerem as informações sobre aves relevantes em processos de análise ecológica, por exemplo.

A experimentação com ferramentas utilizadas no desenvolvimento da Web Semântica realizada no desenvolvimento deste trabalho possibilitou a construção da base de dados semântica sobre a Coleção de Aves do INPA. As informações digitalizadas existentes sobre aves, inicialmente disponíveis somente em planilhas de texto isoladas, foram organizadas, integradas e depois mapeadas para uma base semântica através do Kettle, ferramenta que permitiu o tratamento do dado através de *workflows* e a anotação semântica de forma simples.

Uma vez construída aquela base, o *Rexplorator* foi utilizado para a exploração dos dados de modo a responder às consultas propostas pelos pesquisadores e relacionadas às atividades de pesquisa deles. Essa manipulação foi dirigida por casos de uso que permitiram o desenvolvimento de consultas genéricas e, consequentemente, a sua reutilização por outros pesquisadores.

Uma vez identificadas as consultas comuns propostas pelos pesquisadores, teve-se o subsídio para propôr uma aplicação semântica com a própria ferramenta Rexplorator, possibilitando a utilização da mesma por usuários que não participaram do processo criativo nem da base e nem das consultas, mas que poderão usufruir do conhecimento disponível na Coleção de Aves do INPA.

Outro ponto importante a ser destacado foi conhecer a opinião dos pesquisadores especialistas no domínio de aves sobre as vantagens que o uso das tecnologias desenvolvidas para a WS pode trazer para sua rotina de pesquisa.

5.3.
Trabalhos futuros

Este trabalho propiciou pesquisas adicionais que o Laboratório de Interoperabilidade Semântica (LIS) do INPA está desenvolvendo. Os trabalhos são:

1. A criação de uma aplicação Web semântica para o domínio de aves com a utilização do Synth [Bomfim, 2011] embasada por uma coleção mais completa de informações sobre o domínio e que utilize os dados de outras coleções de aves da Amazônia como os do Museu Paraense Emílio Goeldi-MPEG.
2. Criação de bases semânticas para o domínio de peixes e plantas. Analogamente ao estudo de caso realizado com o domínio de aves, a metodologia proposta por esta dissertação está sendo aplicada em outros domínios das coleções biológicas do INPA, visando a geração de aplicações semânticas para os novos domínios abordados de modo a contribuir na gestão do conhecimento científico dos mesmos.
3. Uma vez criadas bases em outros domínios, comprovação da transdisciplinaridade do conhecimento científico em coleções biológicas através da manipulação dos dados nas bases semânticas.
4. Desenvolvimento de aplicações semânticas para propiciar a aquisição de conhecimento que possa auxiliar os tomadores de decisão do INPA nas atividades relativas à pesquisa e gestão.
5. A publicação de um nodo INPA na LOD como ponto de referência sobre dados relacionados à biodiversidade amazônica.

6
Referências Bibliográficas

ADIDA, B.; BIRBECK, M. **Rdfa primer - bridging the human and data webs**. W3C recommendation, 2008. Disponível em: <http://www.w3.org/TR/xhtml-rdfa-primer/> Acesso em: 3 jan. 2012.

ALBUQUERQUE, A.C.F. **Desenvolvimento de uma Ontologia de Domínio para Modelagem de Biodiversidade**. Manaus, 2011. 147p. Dissertação (Mestrado em Ciência da Computação) – Instituto de Computação, Universidade Federal do Amazonas.

ALBUQUERQUE, A. C. F.; CAMPOS DOS SANTOS, J. **Applying Ontology for Amazon Biodiversity Data Extraction**. In Proceedings of the 9th. World Multi-Conference on Systemics, Cybernetics and Informatics (WSCI 2005). Vol.1; 20050710-13. July 10-13, 2005. Orlando, FL (US).

ALBUQUERQUE, A.C.F.; CAMPOS DOS SANTOS, J.L.; CASTRO JÚNIOR, A.N. **OntoBio: An Ontology for the Amazonian Biodiversity.** To appear at the Proccedings of The International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 4th International Conference on Knowledge Engineering and Ontology Development. Barcelona, Spain, October, 2012.

ANTONIOU, G.; VAN HARMELEN; F. **A Semantic Web Primer**. Second Edition. The Mit Press, 2008.

ARAÚJO, S. F. C. **Explorator: uma ferramenta para exploração de dados RDF baseado em uma interface de manipulação direta**. Rio de Janeiro, 2008. 127p. Dissertação (Mestrado em Informática) – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

AZEVEDO, M. C. **Gerador de aplicações para consultas a bases RDF/RDFS**. Rio de Janeiro, 2010. 119p. Dissertação (Mestrado em Informática) – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

BEACH, J. **Specify Biodiversity Collections Software**. 2010. Disponível em <http://www.specifysoftware.org/Specify/> Acesso em: 15 jan. 2012.

BERNERS-LEE, T.; HENDLER, J. A.; LASSILA, O. **The Semantic Web**. Scientifc American, 2001. Disponível em <http://www.scientificamerican.com/article.cfm?id=the-semantic-web&page=2> Acesso em: 3 jan. 2012.

BERNERS-LEE, T.; HALL, W.; HENDLER, J. A.; O'HARA, K., SHADBOLT, N.; WEITZNER, D.J. **A framework for Web Science**. Foundations and Trends in Web Science, 2006. Disponível em: <http://eprints.aktors.org/594/01/1800000001.pdf> Acesso em: 3 jan. 2012.

BERNERS-LEE, T. **Linked Data**. W3C Design Issues, 2007. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html> Acesso em: 03 jan. 2012.

BERNERS-LEE, T. **Information management a proposal**, 1989. Disponível em: <http://www.w3.org/History/1989/proposal.htm> Acesso em: 03 jan. 2012.

BECKETT, D. **RDF/XML Syntax Specification (Revised)**. W3C Recommendation, 2004. Disponível em: <http://www.w3.org/TR/rdf-syntax-grammar/> Acesso em: 3 jan. 2012.

BECKETT, D.; BERNERS-LEE, T. **Turtle - terse rdf triple language**, 2008. Disponível em: <http://www.w3.org/TeamSubmission/turtle/> Acesso em: 5 jan. 2012.

BIOTA **Programa de Pesquisas em Caracterização, Conservação e Uso Sustentável da Biodiversidade do Estado de São Paulo (BIOTA-FAPESP).** 2010. Disponível em: <http://www.biota.org.br/> Acesso em: 13 jan. 2012.

BLUM, S.; WIECZOREK, J. **TDWG Standard Version 1.4**. July, 2005.

BOMFIM, M. H. S. **Um método e um ambiente para o desenvolvimento de aplicações na Web Semântica**. Rio de Janeiro, 2011. 196p. Dissertação (Mestrado em Informática) – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

BONALDO, A. B.; BRESCOVIT, A. D.; HÖFER, H.; GASNIER, T. R.; LISE, A. A. **A Araneofauna (Arachnida, Araneae) da Reserva Florestal Adolfo Ducke, Manaus, Amazonas, Brasil**. In: Claudio Ruy Vasconcelos da Fonseca; Célio Magalhães; José Albertino Rafael; Elizabeth Franklin. (Org.). "A Fauna de Artrópodes da Reserva Florestal Ducke. Estado atual do conhecimento taxonômico e biológico". 1a. ed. Manaus: INPA, 2009, v. 1, p. 201-222.

BRICKLEY, D.; GUHA, R. V. **RDF Vocabulary Description Language 1.0: RDF Schema**. W3C Recommendation, 2004. Disponível em: <http://www.w3.org/TR/rdf-schema/> Acesso em: 05 jan. 2012.

CAMPOS DOS SANTOS, J. L. **A Biodiversity Information System in an Open Data/Metadatabase Architecture** Ph. D. Thesis. International Institute For Geo-Information Science and Earth Observation. Enschede, The Netherlands, June, 2003. ISBN 90-6164-214-0.

CAMPOS DOS SANTOS, J.L., DE BY, R.A., MAGALHÃES, C.. **A case study of INPA's bio-DB and an approach to provide an open analytical database environment.** International Archives of Photogrammetry and Remote Sensing, 2000, 33 (B4): 155-163.

CHEN, I. A.; MARKOWITZ, V.M. **An overview of the Object Protocol Model (OPM) and the OPM data management tools** . Information Systems, 1995. Disponível em: <www.inf.fu-berlin.de/lehre/SS04/datenbanken/ BioInfDBUnter lagen/objects/objectprotocolModel.pdf> Acesso em: 29 jan. 2012.

CHEN, P. P.S. **The entity-relationship model—toward a unified view of data. ACM Transactions on Database Systems (TODS).** Volume 1 Issue 1, March 1976.

COLWELL, R. K. **Biota, The Biodiversity Database Manager**. Sinauer Associates, 1996.

DELIGIANNIDIS, L., KOCHUT, K. J., SHETH, A. P. **RDF data exploration and visualization**. CIMS'07, em 2007.

ELMASRI, R.; NAVATHE, S. B. **Fundamentals of Database Systems**. Addison Wesley, 6th Edition, 2010. ISBN-10: 0136087209. ISBN-13: 978-0136086208.

FALBO, R. **Experiences in Using a Method for Building Domain Ontologies**. In Proceedings of the Sixteenth International Conference on Software Engineering and Knowledge Engineering, SEKE'2004, pp. 474-477, International Workshop on Ontology In Action, OIA'2004, Banff, Alberta, Canada, June 2004.

FENSEL, D.; FACCA, F. M.; SIMPERL, E.; TOMA, I. **Semantic Web Services**. Springer-Verlag Berlin Heidelberg, 2011. ISBN 978-3-642-19192-3.

FLANAGAN, D.; MATSUMOTO, Y. **The Ruby Programming Languag**e. [S.l.]: O'Reilly Media, 2008.

FOX, P.; HENDLER, J. A. **Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science**. The Fourth Paradigm. Microsoft Research, 2009. p 147-152. 978-0-9825442-0-4.

GT-LINKEDDATABR **LinkedDataBR: Exposição, Compartilhamento e Conexão de Recursos de Dados Abertos na Web (Linked Open Data)**. RT1 - Termo de referência e estado da arte, 2011.

GUIZZARDI, G. **Ontological Foundations for Structural Conceptual Models**. PhD Thesis (CUM LAUDE), University of Twente, The Netherlands. Published as the same name book in Telematica Institut Fundamental Research. Series No. 15, ISBN 90-75176-81-3 ISSN 1388-1795; No. 015; CTIT PhD-thesis, ISSN 1381-3617; No. 05-74. Holanda, 2005.

HALPIN, P. N.; READ, A. J.; BEST, B. D.; HYRENBACH, K. D.; FUJIOKA, E.; COYNE, M. S.; CROWDER, L. B.;FREEMAN, S. A.; SPOERRI, C. **OBISSEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles**. Marine Ecology Progress Series, 316:239-246, 2006.

HEATH, T.; BIZER, C. ***Linked Data: Evolving the Web into a Global Data Space*** (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011. Disponível em: <http://linkeddatabook.com/book> Acesso em: 7 jan. 2012.

HEIM, P.; ZIEGLER, J.; LOHMANN, S. **gFacet: A Browser for the Web of Data**. International Workshop on Interacting with Multimedia Content in the Social Semantic Web. Koblenz, Germany, 2008.

HERMAN, I. **W3C Semantic Web Frequently Asked Questions**. C, 2001. (atualização 2009). Disponível em: <http://www.w3.org/2001/sw/SW-FAQ> Acesso em: 20 mar. 2011.

HULL, R.; KING, R. **Semantic Database Modeling: Survey, Applications, and Research issues**. ACM Computing Surveys, Vol. 19, Nº. 3, September 1987.

INSTITUTO NACIONAL DE PESQUISAS DA AMAZÔNIA (Brasil). **Programa de Coleções e Acervos Científicos (PCAC) Regimento Interno**. Manaus, 05/04/2006. Disponível em: <ftp://ftp.inpa.gov.br/pub/documentos/herbario_inpa/RegimentoPCAC2006.pdf> Acesso em: 05 mar 2012.

INSTITUTO NACIONAL DE PESQUISAS DA AMAZÔNIA (Brasil). **Plano diretor do INPA 2011-2015** / Grupo Gestor de Estratégia do INPA. Manaus : [s.n.], 2011. 46p.

JACOBS, I.; WALSH, N. ***Architecture of the World Wide Web**, Volume One*, 2004. Disponível em: <http://www.w3.org/TR/webarch/> Acesso em: 29 jan 2012.

LEE, B.T.; FISCHETTI, M. **Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor**. Harper, San Francisco, 1999.

MACÊDO, I. T. **Curso de diversidade biológica sobre aves**. Manaus, 2012. Apresentação. Instituto Nacional de Pesquisas da Amazônia.

MACÊDO, I. T.; NAKA, L.N.; COHN-HAFT, M. **Manual de Taxidermia da Coleção de Aves do INPA**. Manual Técnico e Científico – versão preliminar. Manaus, 2011. 30p. Instituto Nacional de Pesquisas da Amazônia.

MACÊDO, J. A. F. **Um modelo conceitual para biologia molecular**. Rio de Janeiro, 2005. 93p. Tese (Doutorado em Informática) - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

MADDISON, D. R.; SCHULZ, K. S. **The Tree of Life Web Project**. Zootaxa, 1668, 2007.

MANOLA, F.; MILLER, E. **RDF Primer**, W3C Recommendation 10 February 2004. Disponível em: <http://www.w3.org/TR/rdf-primer/> Acesso em: 05 jan. 2012.

McCARTNEY, P.; JONES, M. **Using XML-Encoded Metadata as a Basis for Advanced Information Systems for Ecological Research**. Proc. 6th WorldMulticonference Systemics, Cybernetics and Informatics, 7:379-384, 2002.

McGUINNESS, D. L.; VAN HARMELEN, F. **OWL Web Ontology Language Overview**, W3C Recommendation, 2004. Disponível em: <http://www.w3.org/TR/2004/REC-owl-features-20040210/> Acesso em: 10 jan. 2012.

MORRIS, R. A.; STEVENSON, R. D.; HABER, W. **An Architecture for Electronic Field Guides**. J. Intell. Inf. Syst., 29(1):97-110, 2007.

NOY, N. F.; McGUINNESS, D. L. **Ontology Development 101: A Guide to Creating Your First Ontology**. Stanford Knowledge Systems Laboratory Techinical Report KSL-01-05, 2001.

OREN, E., DELBRU, R., GERK, S., HALLER, A., DECKER, S. **ActiveRDF: Objected-Oriented Semantic Web Programming**. WWW2007, Maio de 2007. Banff, Alberta, Canada. Disponível em: <http://www.activerdf.org/> Acesso em: 03 jan. 2011.

PRUD'HOMMEAUX, E.; SEABORNE, A. ***SPARQL Query Language for RDF**, W3C Recommendation*, 2008. Disponível em: <http://www.w3.org/TR/rdf-sparql-query/> Acesso em: 16 jan. 2011.

RAIMOND, Y.; SCOTT, T.; SINCLAIR, P.; MILLER, L.; BETTS, S.; MCNAMARA, F. **Case Study: Use of Semantic Web Technologies on the BBC Web Sites**. United Kingdom, January 2010. Disponível em:

<http://www.w3.org/2001/sw/sweo/public/UseCases/BBC/BBC.pdf> Acesso em: 10 abr. 2012.

REENSKAUG, T. **Thing-model-view-editor: an Example from a planningsystem**. Xerox Parc technical note. 1979.

RUMBAUGH, J. R., BLAHA, M.R., LORENSEN, W., EDDY, F., PREMERLANI, W. **Modelagem e Projetos Baseados em Objetos**. Editora Campos, 2006. ISBN 8535217533.

SAHOO, S., HALB, W., HELLMANN, S., IDEHEN, K., THIBODEAU JR, T., AUER, S., SEQUEDA, J., EZZAT, A.. "A Survey of Current Approaches for Mapping of Relational Databases to RDF". Survey publicado no W3C, janeiro de 2009.

SHADBOLT, N. From Data to D**ecisions: The Power of Information in the Age of the Web of Linked Data. In:Royal Signals Institu**tion Annual Seminar, 19th October, 2010, HQS Wellington, London.

SHAO, K. T.; PENG, C. I.; YEN, E.; LAI, K. C.; WANG, M. C.; LIN, J.; LEE, H.; ALAN, Y.; CHEN, S. Y**. Integration of biodiversity databases in Taiwan and linkage to global databases**. Data Science Journal, pages 2-10, 2007.

TORRES, R. S.; MEDEIROS, C. B.; GONÇALVES, M. A.; FOX, E. A. A **Digital Library Framework for Biodiversity Information Systems**. International Journal on Digital Libraries, 6(1): 3 - 17, February 2006.

UMMINGER, B. L.; YOUNG, S. **Information management for biodiversity: a proposed U.S. National Biodiversity Information Center**. In: REAKA-KUDLA, M.L., WILSON, D.E., WILSON, E.O. (eds.), Biodiversity II: Understanding and Protecting Our Biological Resourses. Washington, D.C., Joseph Henry Press. 1997. p. 491-504.