Linked Data Application for publishing and retrieving data from mammalian collections

Helder M. Cosme[1] [2]
Superintendência de Tecnologia da Informação
Universidade Federal do Rio de Janeiro (UFRJ)[1]
Macaé, Rio de Janeiro, Brazil
heldercosme@macae.ufrj.br

Adriana P. de Medeiros
Instituto de Ciência e Tecnologia
Universidade Federal Fluminense (UFF)[2]
Rio das Ostras, Rio de Janeiro, Brazil
adrianamedeiros@id.uff.br

ABSTRACT

Several research institutions have collections of mammals, which are biological collections that contain important data about the species present in a given region. However, they face problems related to the management of this data and its dissemination, since in many cases this data is stored in spreadsheets, which makes data access and disclosure difficult. This work presents an application that uses the principles of Linked Data and the Darwin Core standard to publish the data of the mammalian collections. It allows accessing data in a friendly way and obtaining more related data, present in other repositories, in order to enrich the data of the samples of species of a collection.

KEYWORDS

Linked Data, Semantic Web, Darwin Core, SHDM, Biodiversity, Collection of mammals.

1 Introdução

As coleções biológicas desempenham papel relevante à saúde pública, agropecuária e demais setores econômicos. A partir da modelagem de dados biológicos relacionados com outros dados ambientais é possível prever o aparecimento e o alastramento de pragas agrícolas, doenças humanas e animais, o que possibilita uma maior eficácia nas ações de combate a epidemias [1].

Diversas instituições de pesquisa, tais como o INPA (Instituto Nacional de Pesquisas da Amazônia) e o NUPEM/UFRJ (Instituto de Biodiversidade e Sustentabilidade), possuem coleções de mamíferos que abrigam dados importantes a respeito das espécies presentes em uma determinada região. Porém, essas instituições geralmente enfrentam problemas relacionados à gerência desses dados e sua disseminação, uma vez que em muitos casos esses dados estão armazenados em planilhas, o que dificulta o acesso aos dados e sua divulgação.

De acordo com Da Rocha, Esteves e Scarano [2] armazenar dados em planilhas deixa os mesmos sujeitos à degradação devido a problemas como: desvinculação de pesquisadores do projeto; morte de pesquisadores; e problemas com hardware ou software que inviabilizam o acesso aos dados, muitas vezes sem possibilidades de recuperação. Também dificulta a disseminação e a vinculação desses dados com outras bases.

Devido à necessidade das instituições de divulgar suas coleções, se faz necessária à utilização de tecnologias que possibilitem o acesso aos dados. É importante, também, que os dados da coleção sejam relacionados com outras bases, a fim de melhorar o entendimento sobre as espécies.

Segundo Bizer, Heath e Berners-Lee [3], *Linked Data* tem sido utilizado como meio para a publicação de dados de forma a facilitar o acesso aos dados de uma instituição. Ele refere-se à publicação de dados na web de forma que os mesmos possam ser processados por agentes de software. *Linked Data* foi proposto com o objetivo de ligar dados na web, podendo esses estar em bases de dados de diferentes organizações em diferentes locais geográficos. Assim, os recursos de *Linked Data* podem ser utilizados para relacionar as informações sobre coleções de mamíferos aos dados de outras bases, com o intuito de buscar informações sobre mamíferos que venham a complementar os dados de espécies constantes nessas coleções.

Este artigo apresenta uma aplicação *Linked Open Data* denominada Mamíferos, com o objetivo de aprimorar o acesso aos dados de coleções de mamíferos, padronizar a inserção desses dados e melhorar a gerência dos mesmos. A primeira versão dessa aplicação foi desenvolvida para a coleção de mamíferos do NUPEM/UFRJ.

2 Referencial teórico

Coleções biológicas são centros depositários de material biológico. As coleções abrigam não só os espécimes coletados e estudados, mas também as informações associadas aos indivíduos e às populações de cada espécie [1]. Segundo Veiga [4], o processo de produção de dados de espécies é composto por duas etapas. A primeira consiste em coletar o organismo na natureza e registrar em meios digitais ou manuscritos. A segunda está relacionada à digitalização das informações previamente selecionadas e interpretadas.

A maioria dos espécimes constantes nas coleções de mamíferos é oriunda de coletas feitas em determinados pontos de coleta. Essa atividade de coleta é denominada Evento de Coleta, que é a descrição de uma ação que ocorre em determinado local e tempo. Em um evento de coleta são registrados dados de

In: XVIII Workshop de Ferramentas e Aplicações (WFA 2019), Rio de Janeiro, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2019.

ISSN: 2596-1683

identificação do animal como nome da espécie, sexo, dimensões do corpo e peso. Também são armazenados dados geográficos referentes à coleta.

Para facilitar o intercâmbio dos dados dessas coleções é utilizado o padrão *Darwin Core* [5] para a estruturação dos dados das espécies presentes nas coleções de mamíferos. *Darwin Core* é um padrão amplamente utilizado para publicação e integração de dados sobre biodiversidade. Ele possui um vocabulário RDF (*Resource Description Framework*) [6] que auxilia na estruturação dos dados no contexto de *Linked Data*.

Na modelagem da aplicação Mamíferos foi utilizado o método SHDM (*Semantic Hypermedia Design Method*) [7]. SHDM é uma evolução do OOHDM (*Object-Oriented Hypermedia Design Method*) [8], um método que utiliza técnicas de orientação a objetos para o projeto de aplicações hipermídia. Ele mantém a mesma estrutura de fases de modelagem do OOHDM. No entanto, enriquece os modelos criados, através de construções com maior poder expressivo e formaliza alguns novos conceitos, como a estrutura de acesso facetada, levando em consideração os formalismos e primitivas introduzidas pela Web Semântica. O SHDM é composto por cinco etapas: levantamento de requisitos, que identifica os atores e tarefas a serem apoiadas pela aplicação; projeto conceitual, onde são criados o Esquema conceitual SHDM, uma Ontologia conceitual SHDM e um conjunto de instâncias; projeto navegacional, no qual são criados o Esquema de Classes Navegacionais e o Esquema de Contextos Navegacionais; projeto de interface abstrata; e implementação [7].

3 Arquitetura da Aplicação Mamíferos

A arquitetura definida para a aplicação possui três camadas: camada de apresentação, de aplicação e de dados, como mostra a Figura 1. A camada de apresentação é responsável por prover mecanismos para o usuário ter acesso à aplicação. A camada de aplicação é responsável por obter os dados vindos da camada de apresentação e estruturá-los semanticamente através da utilização dos termos do vocabulário *Darwin Core* em RDF. Essa camada também é responsável por efetuar as consultas aos *datasets* da camada de dados através de consultas SPARQL (*SPARQL Protocol and RDF Query Language*) [9] e disponibilizar uma resposta útil para a camada de apresentação. Para isso ela utiliza os vocabulários das ontologias do *DBpedia* [10] e do *Geonames* [11]. A camada de aplicação, além de estruturar os dados semanticamente, também é responsável por fazer a ligação de recursos presentes no *dataset* local com outros recursos, presentes nos *datasets* do *DBpedia* e do *Geonames*. Essa ligação de dados possibilita a obtenção de informações que enriquecem os dados armazenados localmente.

A camada de dados abriga as triplas RDF do *dataset* da coleção de mamíferos. As triplas são criadas pelo componente *Linked Data* que faz a estruturação dos dados e os insere no *triple store*. Ela também é responsável por prover e gerenciar o acesso ao *dataset* local. Essa camada também permite que agentes de software possam obter os dados da base local através do SPARQL *Endpoint* disponibilizado por essa base. Assim, é possível que outras aplicações possam acessar os dados na base da coleção de mamíferos local bem como vincular seus dados a eles.

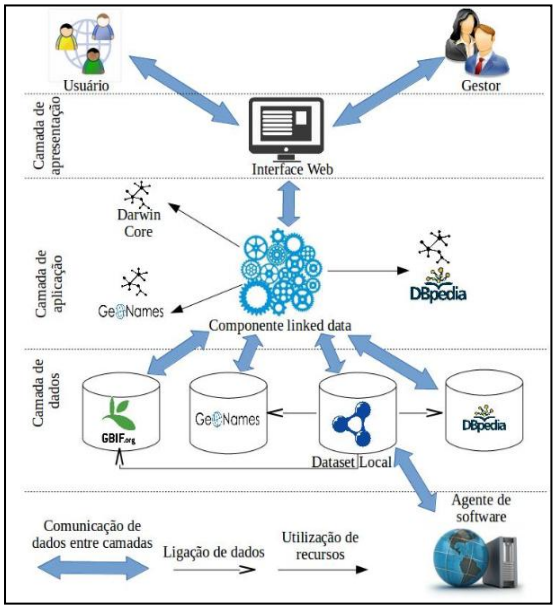

Figura 1: **Arquitetura da aplicação Mamíferos**

A estrutura de navegação da aplicação Mamíferos, representada com o esquema de contextos de navegação do método SHDM, é ilustrada na Figura 2.

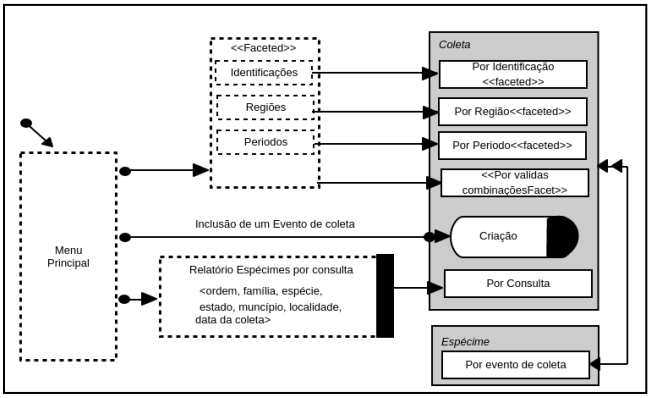

Figura 2: Esquema de contextos navegacional

Esse diagrama mostra os índices (retângulos tracejados) e contextos (conjuntos de objetos, retângulos em cinza) que podem ser explorados pelo usuário. A partir do índice "Menu Principal" é possível acessar a funcionalidade de inclusão de um evento de coleta e os índices referentes às diferentes consultas de espécimes. Ao selecionar um item em um índice, o usuário tem acesso aos dados de um conjunto de coletas, por exemplo, coletas de uma determinada região (contexto "Coleta por Região"). A partir de uma coleta também é possível navegar pelos espécimes dessa coleta (contexto Espécime por Evento de Coleta).

4 A aplicação Mamíferos

A aplicação foi modelada usando o método SHDM [7] e desenvolvida na linguagem Java[1], com a tecnologia JSF (*Java Server Faces*)[2] combinada com o *framework Primefaces*[3], a base *RDF Allegrograph*[4] e o *framework Jena*[5]. A base de dados RDF foi criada seguindo o guia RDF e o padrão *Darwin Core*. A aplicação possui uma licença do tipo GNU *General Public License* (GPL).

A Figura 3 mostra a página inicial da aplicação, que pode ser acessada pelo público em geral. Nela são apresentadas as coletas de espécie mais recentes, presentes na coleção, e as opções de menu que permitem acessar as funcionalidades da aplicação.

Figura 3: Página inicial da aplicação Mamíferos

4.1 Cadastro de dados de espécimes na coleção

Essa funcionalidade é acessada pela opção "Inserir Coleta" e é responsável por estruturar os dados dos espécimes em RDF seguindo o padrão Darwin Core, fazer as ligações semânticas com as bases *DBpedia* e *Geonames*, e inserir os dados na base RDF local. Esse cadastro é dividido em duas etapas, onde a primeira é responsável por inserir os dados referentes à coleta, como local, identificação da espécie, data da coleta. Já a seg 1) unda é responsável por inserir dados relacionados ao espécime coletado, como peso, sexo, comprimento.

Inicialmente a aplicação apresenta um formulário, na aba Dados da Coleta ilustrada na Figura 4, para a inserção dos dados de coleta na base RDF. Após o preenchimento e selecionada a opção de "salvar", a aplicação cria uma URI (Universal Resource Identifier) para identificar essa coleta unicamente, de modo que ela possa ser acessada via consulta SPARQL diretamente. Essa coleta é estruturada como uma instância da classe RDF Event do padrão Darwin Core. A técnica de coleta, coletor, data da coleta e observações, relacionados à coleta, são estruturadas usando os predicados dwc:samplingProtocol, dwc:recordedBy, dwc:eventDate, dwc:eventRemarks, respectivamente.

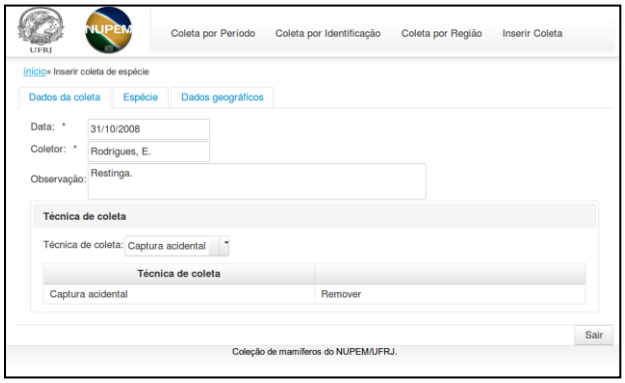

Figura 4: Página de cadastro de dados de coleta de espécie

A aba Espécie permite inserir dados sobre a identificação da espécie como nome científico, ordem e família. Para facilitar o preenchimento, são carregados do *DBpedia* dados sobre a identificação de mamíferos. Assim, o usuário pode selecionar a ordem, a família e o nome científico que foi carregado do *DBpedia*. No momento da inserção, somente o link RDF para essas informações fica presente na base local da coleção.

Os dados de identificação da espécie são estruturados usando a classe *dwc:Identification*, como mostra a Figura 5. Para identificar a amostra de espécie, em vez de estruturar cada dado, é feito um link RDF para a base do *DBpedia*, que contém os dados de identificação relacionados a essa amostra de espécie. Esse link é feito através de uma consulta SPARQL, que utiliza o nome científico da amostra de espécie para obter a URI, referente a essa espécie, na base do *DBpedia*. Esse vínculo permite a descoberta de mais informações sobre essa espécie como, descrição resumida, imagem e classificação taxonômica. Esse vínculo com o *DBpedia* é feito com o predicado *dwciri:toTaxon* do vocabulário *Darwin Core*. Também é feito um link RDF para o portal GBIF (*Global Biodiversity Information Facility*) [12], com o objetivo de ligar os dados de identificação dessa coleção aos dados de identificação que constam nesse repositório.

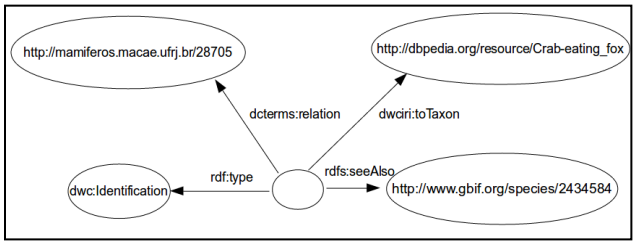

Figura 5: Dados de identificação da espécie

[1] https://www.oracle.com/java/
[2] http://www.oracle.com/technetwork/java/javaee/overview-140548.html
[3] https://www.primefaces.org/
[4] https://franz.com/agraph/allegrograph/
[5] https://jena.apache.org/

A aba Dados geográficos (Figura 4) permite a inserção de dados geográficos a respeito da coleta da espécie. Para facilitar o preenchimento, são carregados do *Geonames* os dados dos estados e municípios. Assim, o usuário pode selecionar o estado e o município referente à coleta. No momento da inserção, na base local, serão armazenados os dados de observação sobre o local da coleta, localidade, latitude, longitude e o link RDF que aponta para os dados que estão na base do *Geonames*.

Os dados geográficos relacionados à coleta são estruturados pela aplicação usando a classe *dcterms:Location*. Dados como localidade, latitude, longitude e observação sobre o local da coleta foram respectivamente estruturados com os predicados RDF, *dwc:municipality*, *dwc:decimalLatitude*, *dwc:locationRemarks*. Já o predicado *dwciri:inDescribedPlace* estruturou o link RDF feito para a base do *Geonames*, com o objetivo de obter informações geográficas sobre a coleta como estado e município. Essa ligação também proporciona qualidade aos dados, pois eles não são digitados pelo usuário e sim selecionados a partir dos dados obtidos do Geonames, e vinculados com essa base de dados utilizando *Linked Data*.

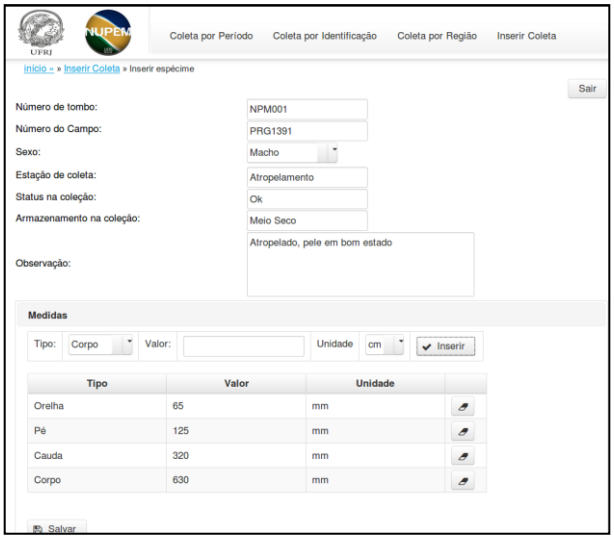

Figura 6: Cadastro de dados de amostras de espécie

Após o cadastro, os dados da coleta são registrados na base RDF local e os links semânticos com as bases *DBpedia*, *Geonames* e *GBIF* são estabelecidos. Em seguida, a página de inserção de dados de espécimes é disponibilizada, como mostra a Figura 6.

Os dados do espécime são registrados pela aplicação como instâncias da classe *dwc:Occurrence*. Para identificar essa amostra de espécie é criada uma URI. Esses dados estão relacionados diretamente à amostra da espécie coletada. Na coleção são catalogados os seguintes dados relacionados às espécimes: número do campo, número do tombo, sexo, estação de coleta, status na coleção, armazenamento na coleção e observação.

A coleção também conta com dados de medidas relacionadas a cada espécime. Esses dados são estruturados com a classe *dwc:MeasurementOrFact*. Na coleção são catalogados os seguintes dados sobre medidas: tipo, valor e unidade de medida.

4.2 Consulta aos dados da coleção

A aplicação possui três tipos de consultas: busca de coleta por período, por identificação e por região. A Figura 7 mostra a página de busca por período e seu resultado.

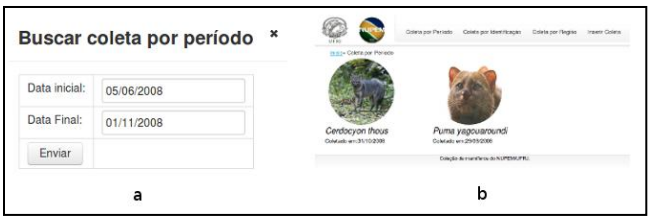

Figura 7: Busca por Período (a) e página com o Resultado da busca de coleta por período (b)

Após a execução da busca a aplicação exibirá as espécies coletadas nesse período, como mostra a Figura 7 - b. Nesse exemplo a busca retornou duas espécies. Ao clicar sobre o nome da espécie, a aplicação exibe todas as informações sobre a espécie, como mostra a Figura 8. Nessa página são apresentadas uma descrição sobre a espécie, uma imagem e seu nome científico. Estas informações são extraídas da base *DBpedia*. Ela também mostra detalhes da coleta como local e dados específicos da coleta da espécie. A opção "Saiba mais sobre essa espécie", direciona o usuário ao portal GBIF, através do link RDF estabelecido dinamicamente no momento do cadastro da espécie.

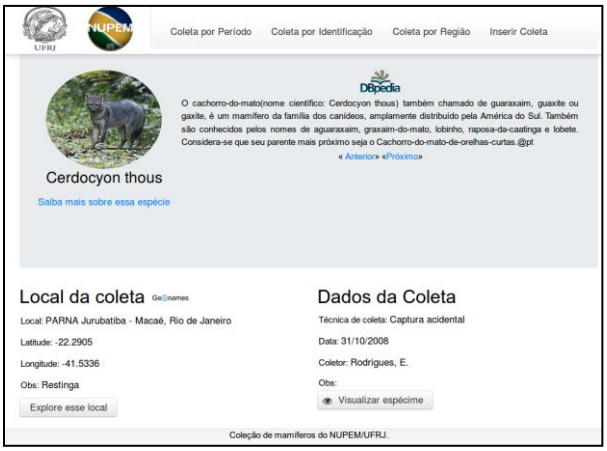

Figura 8: Página com informações sobre a espécie

Também é possível obter mais informações sobre o local da coleta através da opção "Explore esse local". Ela disponibiliza um mapa do local onde o espécime foi coletado, possibilitando a visualização do local onde ocorreu a coleta. As informações de localidade são obtidas através do link RDF feito para a base do *Geonames*. Finalizando, é possível acessar os dados dos espécimes, relacionados à espécie, através do botão "Visualizar

espécime". Nessa página também é possível navegar pelas informações das espécies retornadas na consulta, usando as âncoras "Próximo" e "Anterior", localizadas abaixo da descrição da espécie.

5 Trabalhos relacionados

Em BioDSL [13] foi proposta uma linguagem específica de domínio para converter os dados de biodiversidade do formato CSV para o RDF. Além disso, BioDSL permite ligar esses dados, enriquecendo-os com dados provenientes de outros repositórios RDF, e criar padrões para a geração de URIs para cada recurso.

Em [14] foi proposto um *framework* para gerenciar, tratar e integrar dados ecológicos referentes ao projeto de Pesquisas Ecológicas de Longa Duração (PELD) realizado na baía de Guanabara. *Linked Data* foi utilizado para proporcionar integração entre as bases de dados desse projeto de pesquisa.

Em Theophrastus [15] foi proposto o uso de *Linked Data* em uma aplicação que explora páginas web ou documentos buscando informações sobre espécies. Ao encontrar, ela marca essa informação no documento ou página web e mostra, através de links semânticos, mais informações sobre a espécie.

Semantic Web Interactive Gazetteer (SWI) [16] usa Linked Data para ligar seus dados com outros dados na web e armazená-los em RDF. Ele também provê um mecanismo de acesso a esses dados através de *Endpoints* GeoSPARQL. SWI utiliza as bases *Geonames* e *DBpedia* para tratar os dados e verificar se eles estão corretos, evitando a inserção de dados incorretos em sua base.

Todos os trabalhos encontrados utilizam os princípios de *Linked Data* para resolver problemas referentes aos dados de pesquisa em Biodiversidade. No entanto, esses trabalhos não tratam dados sobre coleções de mamíferos, foco deste trabalho.

6 Conclusões

Este trabalho apresenta uma solução para apoiar a publicação e integração de dados sobre coleções de mamíferos usando *Linked Data*, visando facilitar o acesso aos dados da coleção não só por pesquisadores, mas também pelo público em geral.

As contribuições deste trabalho incluem a estruturação dos dados das espécies, presentes na coleção, seguindo o modelo RDF do padrão *Darwin Core*, a criação de um repositório RDF com os dados reais dessa coleção e a publicação desses dados na web utilizando os princípios de *Linked Data*.

Com a estruturação dos dados da coleção seguindo o guia RDF do padrão *Darwin Core* e a aplicação Mamíferos, os dados de coleções de mamíferos podem ser organizados de forma padrão e disponibilizados para processamento por máquinas, através de um SPARQL *EndPoint*, o que proporciona um acesso facilitado por soluções computacionais e uma possível integração dessas soluções.

A aplicação Mamíferos passou por uma avaliação inicial, realizada em formato de teste por dois professores e cinco alunos do NUPEM. O teste foi dividido em duas etapas. A primeira foi referente ao cadastro de uma coleta de espécie e de uma amostra dessa espécie. Na segunda, os mecanismos de busca presentes na aplicação foram utilizados para consultar os dados cadastrados. Nessa avaliação foi utilizado um questionário contendo seis questões usando como base a escala Likert [17]. As questões foram formuladas para avaliar o suporte dado pela aplicação aos pesquisadores no registro e no acesso aos dados das espécies da coleção de mamíferos do NUPEM. Os resultados iniciais indicaram que a aplicação facilita o cadastro, a consulta, a navegação e a visualização de uma amostra de espécie presente na coleção, além de contribuir para uma melhor compreensão da espécie e local da coleta.

Um trabalho futuro previsto é a implementação de uma funcionalidade de consulta facetada do SHDM, que permitirá combinar as consultas por data, identificação e local da coleta. Outro trabalho futuro é a finalização do módulo de gestão de dados da coleção e controle de acesso de usuários para possibilitar aos curadores das coleções cadastrarem novos usuários e fazer um controle das modificações nos dados, como quem inseriu ou alterou algum dado. Também será implementado um relatório a respeito dos dados da coleção. Por fim, uma nova avaliação da aplicação também será realizada.

REFERÊNCIAS

[1] PEIXOTO, A. L. et al. 2006. Diretrizes e estratégias para a modernização de coleções biológicas brasileiras e a consolidação de sistemas integrados de informação sobre biodiversidade. Brasília: Centro de Gestão e Estudos Estratégicos: Ministério da Ciência e Tecnologia, p. 145–182.

[2] DA ROCHA, C. F. D.; ESTEVES, F. DE A.; SCARANO, F. R. 2004. Pesquisas de longa duração na Restinga de Jurubatiba: ecologia, história natural e conservação. São Carlos, SP. RiMa Editora.

[3] BIZER, C.; HEATH, T.; BERNERS-LEE, T. 2009. Linked data-the story so far. International journal on Semantic Web and Information Systems. v. 5, n. 3, p. 1–22.

[4] VEIGA, A. J. 2012. Um estudo sobre qualidade de dados em biodiversidade: aplicação a um sistema de digitalização de ocorrências de espécies. 101 f. Dissertação (Mestre em Ciências). Escola Politécnica da Universidade de São Paulo (USP), Departamento de Engenharia de Computação. São Paulo, SP.

[5] WIECZOREK, J. et al. 2012. Darwin core: An evolving community-developed biodiversity data standard. PLoS ONE, v. 7, n. 1, p. e29715, 6 jan.

[6] CYGANIAK, R., LAANTHALER, M., WOOD, D. Resource Description Framework (RDF): Concepts and Abstract Syntax. Disponível em: <https://www.w3.org/TR/rdf11-concepts/>. Acesso em: 29 jun 2019.

[7] LIMA, F. 2003. Modelagem semântica de aplicações na WWW. Tese (Doutorado em Informática) - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro PUC-Rio, Rio de janeiro, RJ.

[8] SCHWABE, D.; ROSSI, G. 1998. An Object Oriented Approach to Web-based Applications Design. Theory and Practice of Object Systems, v. 4, n. 4, p. 207.

[9] PRUD'HOMMEAUX, E.; SEABORNE, A. SPARQL Query Language for RDF. Disponível em: <http://www.w3.org/TR/rdf-sparql-query/>. Acesso em: 29 jun 2019.

[10] AUER, S. et al. 2007. DBpedia: A Nucleus for a Web of Open Data. The semantic web. Springer.

[11] WICK, M.; VATANT, B. 2012. The Geonames geographical database. Disponível em: <http://www.geonames.org>. Acesso em: 29 jun 2019.

[12] GBIF. 2012. Global Biodiversity Information Facility (GBIF). Natural History, v. 29, n. March, p. 1–2.

[13] SERIQUE, K. J. DO A. et al. 2016. BioDSL: a domain-specific language for mapping and dissemination of biodiversity data in the LOD. Congresso da Sociedade Brasileira de Computação, XXXVI; Brazilian e-Science Workshop.

[14] MOURA, A. M. D. C. et al. 2012. Integrating Ecological Data Using Linked Data Principles. Joint V Seminar on Ontology Research in Brazil, p. 156–167.

[15] FAFALIOS, P.; PAPADAKOS, P. 2014. Theophrastus: On demand and real-time automatic annotation and exploration of (web) documents using open linked data. Journal of Web Semantics, v. 29, p. 31–38.

[16] CARDOSO, S. D. et al. 2016. SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data. Future Generation Computer Systems, v. 54, p. 389–398.

[17] LIKERT, R. 1932. A technique for the measurement of attitudes. Archives of psychology, New York, Columbia University Press.