BLO: Batata Lake (Oriximiná/PA) Application Ontology

Adriano Neves de Souza, Adriana Pereira de Medeiros

Instituto de Ciência e Tecnologia – Universidade Federal Fluminense – Rio das Ostras
Rio de Janeiro – RJ – Brazil

adriano_souza@id.uff.br, adrianamedeiros@puro.uff.br

Abstract. *This work presents the BLO ontology (Batata Lake Ontology), an application ontology that describes in a structured way the data of research done by limnology researchers of Federal University of Rio de Janeiro (UFRJ) Macaé in Batata Lake (Oriximiná/PA). The main contribution of the BLO is the creation of a research data repository in RDF and the BLS application (Batata Lake System), a semantic web application to support researchers in environmental impact assessments, in preservation areas settings, in species protection and recovery of degraded areas, among other activities.*

1. Introduction

The ecological complexity of aquatic ecosystems caused by the large volume of sampling data creates difficulties to understand the environment and species, as well as the relationship between them. This understanding generates scientific data and knowledge, which provides recovery alternatives or mitigation of external impacts in the ecosystem [Bozelli et al. 2000]. Governments and organizations are encouraging solutions to share the knowledge of ecology. For example, the PELD (Long Term Ecological Program) [Esteves et al. 2004] was created by the Brazilian government to encourage the organization of research data on ecosystems. Limnology researchers of the UFRJ Macaé-RJ have been working for decades in research about the Batata Lake, an Amazonian aquatic ecosystem, located at Oriximiná-PA, that suffered environmental impacts due to the tailings generated by bauxite production [Bozelli et al. 2000]. This lake has been monitored and studied since the 80's in order to obtain knowledge of its ecosystem and mitigate these impacts. The lack of structuring and formalization of the large volume of generated data makes their analysis difficult, and limits the scope of the researchers in the search for new knowledge.

The application of Semantic Web technologies for the management and understanding of research data has been widely discussed currently. The ontologies usage in biodiversity has been appointed as a solution for obtaining scientific knowledge [Campos et al. 2011]. Ontologies for biodiversity are presented in [Moura et al. 2012], [Campos et al 2011] and [Amanqui et al 2013], but they do not describe terms proposed in this work.

This paper presents the application ontology BLO (Batata Lake Ontology) that describes the data of analysis and samplings obtained by limnology researchers of UFRJ Macaé-RJ in order to support their researches. It also presents the BLS web application for supporting the lake recovery analysis and the search for solutions that mitigate the environmental impacts. An exploratory study performed to validate the ontology is presented. Then, some conclusions and future works are discussed.

2. Batata Lake Application Ontology

Application ontologies describe concepts of a domain and specific tasks for implementing systems, the practical part [Guarino, 1997]. BLO was created following the Ontology Development 101 [Noy et al, 2001] guide. It was specified with the OWL (Ontology Web Language), specifically OWL DL 2, with 35 classes and 222 axioms. The domain was defined as Batata Lake. Thus, the ontology will be used to support the limnology researches of UFRJ Macaé, organizing research data, providing relevant information to the environmental impacts mitigation in this lake and preparing these data for online publication when needed. The ontology scope was determined by drafting the following competency questions: i) What is the sample period with the highest concentration of chlorophyll in a given year? ii) What flood pulse had the highest amount of turbidity in a given year? iii) What flood pulse had the highest percentage of organic matter in the sediment in a given year? iv) What is the flood pulse of a certain period? v) What samplings were done in impacted areas in a given period?

Searches were performed in the ontology repositories *DAML Ontology Library* (www.daml.org/ontologies/), *Protégé_Ontology_Library* (protegewiki.stanford.edu/wiki/Protege_Ontology_Library), *Schemapedia* (datahub.io/pt_BR/dataset/schemapedia) and *Swoogle* (swoogle.umbc.edu/), in order to find ontologies related to this work. The ontologies HydroBodyOfWater (sweet.jpl.nasa.gov/2.0/hydroBodyOfWater.owl) and Geography (www.daml.org/ontologies/412) contain some generic terms with descriptive features related to the proposed ontology, but they do not address the domain of this work. After the BLO definition, Albuquerque et al (2015) proposed sub-ontologies as complements to biodiversity ontology OntoBio to create a fieldwork sample vocabulary. The reuse of this vocabulary in the BLO ontology is a future work.

Figure 1 shows the graph preview of the main classes of the BLO. The vertices are classes or concepts defined in the ontology. The edges, which have a one direction, are the relations between classes, also called object properties. The *Sampling* class describes the collected sample by the researcher in the sampling stations, represented by *SamplingStation* class. *SamplingStation* has two data properties: *coordenates* and *impacted*, which respectively specify the geographical location of the sampling station and whether it is an impacted area or not. The object property *isDoneOn* determines the relation between *Sampling* and *SamplingStation*. The relation *isDoneDuring* between *Sampling* and *Period* expresses that a sampling is done in a particular period. The number of possible relations is limited by the amount of sampling stations that had some collected sample. The *FloodPulse* class specifies the lake flood pulses, which are the process stages of filling and emptying of the lake. This class has no data property, because the identification of instances is done by the URI (Flood, HighWater, ebby, LowWater). The *Period* class contains the data property *date* that describes the month and year in which the sampling is done. It is related to *FloodPulse* class by the object property *determines*. This property describes the relation between the months of the year and the flood stages of the lake, which can suffer changes over the years, because there is no standard in the establishment that a month will have a particular flood pulse. The *Sediment* and *Water* classes represent all data collected of sediment and water in the lake and they are related to the sampling by the object property *isSampliedBy*. All sampling data related to water are described by data properties of the classes *Water*, *SuspendedMatterial*, *Aluminum*, *Chorophyll*, *Iron*, *Nitrongen*, *Oxigen* and *Phosphor*.

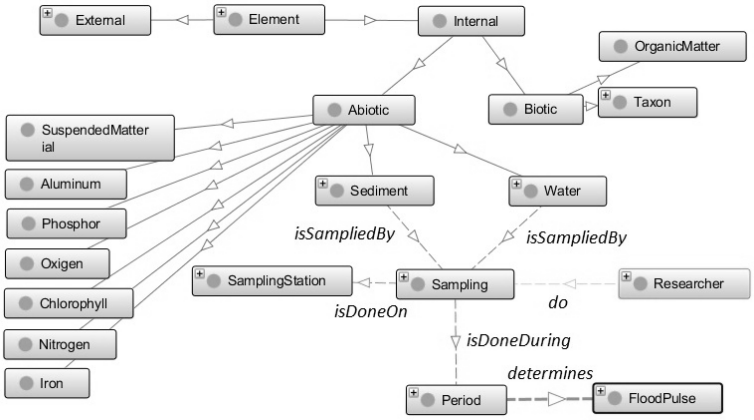

Figure 1- BLO Classes and Properties (partial)

The object property *isDoneOn* between *Sampling* and *SamplinStation* is defined with the restriction *FunctionalProperty*. Thus, a sampling *x* can be done in only one sampling station *y*. Using the triple *Sampling-> isDoneOn-> SamplingStation* is possible seek sampling information grouped by sampling stations. The object property *determines* is defined as inverse of *isDeterminedBy*. It allows that when answered the competency question "What is the flood pulse of a certain period?", the *reasoner* identifies the inverse relation *isDeterminedBy* and retrieve any instance that has the inverse as relation. Restrictions like these add semantic details to the data model and with *reasoners* the queries can obtain more accurate results, as shown in the section 3.

The BLO instances were obtained from actual research data of the Batata Lake stored in the last 26 years in spreadsheets. These data were automatically exported to RDF [Graham; Jeremy, 2004] using the BLO vocabulary and stored in a repository using the AllegroGraph 4.14 (http://franz.com/agraph/).

3. BLS Web Application

BLS (Batata Lake System) was developed to provide accurate information of the lake for researcher analysis. It was implemented in JAVA with JENA library (http://www.w3.org/2001/sw/wiki/Jena), which allows connecting the application to the RDF repository. JENA is a Java framework for building Semantic Web applications and has support for manipulating RDF triples, OWL, SPARQL [Eric; Andy, 2008] queries and includes an inference engine (*Reasoner*). The BLS interface was developed in Portuguese. Figure 2 presents the Period query page, which allows searching a given period by date (Período) or flood pulse (Pulso de Inundação). All periods of the selected pulse are raised when the page is submitted. During query performing, the application accesses the stored data in the RDF repository and run the query in SPARQL. Frame 1 presents the SPARQL query executed from page shown in Figure 2 and answers the competency question "What is the flood pulse of a certain period?". Thus, the BLS application displays the query result illustrated in the Figure 2, which shows that the flood pulse was Low Waters (AguasBaixas). Note that the data can be described using the relation *isDeterminedBy* in the RDF repository instead of *determines*. However, the query result would be the same, because these properties were defined as inverse in the BLO. The “eye” icon displays all requested period data, but the result will not be presented here due to space limitations.

Figure 2- Period Query

Figure 3 - Sampling Query

```
select distinct ?date ?floodpulse
    WHERE {
            ?period  rdfs:type  blo:Period.
            ?period  blo:determines ?floodpulse.
            ?period  blo:samplingDate  ?date.
            FILTER (?date = "12/2001")
           }
```

Frame 1 – Period SPARQL query

```
select  ?date ?station
    WHERE {
            ?period   a  blo:Period.
            ?sampling  a  blo:Sampling.
            ?station   a  blo:SamplingStation.
            ?sampling blo:isDoneOn ?station.
            ?station   blo:impacted ?impacted.
            ?sampling  blo:isDoneDuring ?period.
            ?period       blo:samplingDate  ?date.
            FILTER (?date = "12/2001").
            FILTER (?impacted = "true")
            }
  ORDER BY (?date)
```

Frame 2 – Sampling SPARQL Query

The sampling query page presented in Figure 3 allows searching the samplings done in a period or by a particular researcher in impacted area or not. It answers the competency question "Which samplings were done in impacted areas in a given period?". The application can consider the filter by researcher, otherwise it will be considered by the period. The samplings can be selected by sampling stations. The FILTER term in Frame 2 is used to determine the sampling period and the sampling station type that the researcher wants to get as answer in the sampling query page. The query result helps to evaluate the samplings which were done in impacted areas and thus comparing with samples done in non-impacted areas, in order to historically evaluate the behavior and recovery of the environment.

4. Exploratory Study

In order to evaluate the data model defined by BLO and the BLS application, a small exploratory study was conducted. The hypothesis was that the use of Semantic Web technologies for describing the Batata Lake data would facilitate the access and analysis of these data. The study was performed from a test divided into two stages: the execution of a search activity using the BLS application and the fill of an evaluation questionnaire. The activity was evaluating the water turbidity of a sample in a given period, considering as parameter the sampling data of non-impacted areas done in the same period. This is important for the researchers, since that allows evaluating the progress of the lake recovery. The study involved seven participants. The choice of them was premised on the experience and engagement with lake researches. Two of the participants, one PhD researcher and one master student, accompanied and provided all the necessary for understanding the domain and definition of competency questions.

The goal of the study was to evaluate how the research data started to be searched and analyzed using the BLS. It was not stipulated time for performing the activity. At the end of the activity each participant filled a joint questionnaire with the following questions: 1) *Do the searches available in the web application allow finding and relating the data of the samplings? Why?* 2) *Do the results obtained by the searches facilitate the comparison of the data and the analysis of the lake recovery? Why?* 3) *Would you use this application again to query and analyze your research data? Why?* 4) *Do the terms and system's menu options correspond to the everyday reality of research about the lake? If the answer is no, list the terms that do not match the reality.* 5) *Considering a scale of one to five, with option 1 equal bad and 5 equal great, how do you rate the form of searching available in the web application, comparing it with that currently performed in Excel spreadsheets?*

Most participants (five of them) answered "yes" to the questions and valued the new way to query research data. Six participants said that would use the BLS application again, as this tool significantly reduces the time spent looking for a data, enabling faster analysis. Six of them said that the vocabulary was defined according to the everyday reality of research about the lake. This indicates that the BLO ontology was well defined according to the domain. The test results also allowed identifying problems and difficulties in finding and analyzing the data. In the issue 2, the answers of four participants indicated that the queries results did not facilitate the data comparison and the lake recovery analysis, because the way the results were presented. They informed the search filter by period should be only for year interval with a flood pulse filter to facilitate analysis based on different periods and years. In addition, they suggested the choice of some variables, such as turbidity or chlorophyll, presented in parallel all the values separated by sampling stations, impacted or not. It would allow analyzing a historical series of data and effectively evaluate the lake recovery. Plus, they observed that the application navigability would be more intuitive with the access to the samplings from the data of a given period.

5. Conclusion and Future Work

This paper presented the BLO ontology for semantically describing data of the research done by limnology researchers of UFRJ-Macaé on Batata Lake. The semantic description of these data enables richer queries about the lake through inferences done by *reasoners*. In addition, it provides a vocabulary of common terms used in other researches about the Batata Lake. The main contributions of this ontology is the creation of a research data repository in RDF and the development of the BLS system, a semantic web application to support researchers to query and analysis the research data about this lake. The aim is supporting the production of scientific knowledge from the analysis made by semantic queries and preparing the data for online publication when needed. An initial exploratory study was done to validate the ontology and the application. The tests showed the BLO relevance and quality and some necessary changes in the BLS application. After implementing these changes, a new experiment will be conducted to validate them.

A future work is using the ontology proposed by Moura et al (2012) and the BLO ontology for describing the species existing in the Batata Lake. Another future work is sharing BLO so that other researchers that study this lake can use it to support their research. Moreover, some terms related to fieldwork sampling context of the

OntoBio [Albuquerque et al, 2015] can be reused.

References

ALBUQUERQUE, A. C. F., CAMPOS DOS SANTOS, J. L., DE CASTRO JÚNIOR, A. N. OntoBio: A Biodiversity Domain Ontology for Amazonian Biological Collected Objects. 48th Hawaii International Conference on System Sciences, p. 10, 2015.

AMANQUI, F. K. M.; SERIQUE, K. J.; LAMPING, F.; CAMPOS, J. L.; ALBUQUERQUE, A. C. F.; MOREIRA, D. A. Implementing an Architecture for Semantic Search Systems for Retrieving Information in Biodiversity Repositories. Simpósio Brasileiro de Banco de Dados, p. 1–6, 2013.

BOZELLI, REINALDO L.; ESTEVES, FRANCISCO A.; ROLAND, F. Lago Batata: Impacto e Recuperação de um Ecossistema Amazônico. UFRJ/SBL- RJ, 2000.

CAMPOS, J. L.; NETTO, J. F. D. M.; CASTRO, A. N. DE; ALBUQUERQUE, A. C. F. Ontologias para Interoperabilidade de Modelos e Sistemas de Informação de Biodiversidade, 2011.

ERIC, P.; ANDY, S. 2008 “SPARQL Query Language for RDF”. W3C. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

ESTEVES, F. A. ; SCARANO, F. R. ; ROCHA, C. F. D. Pesquisa de Longa Duração na Restinga de Jurubatiba: Ecologia, História Natural e Conservação. 1. ed. Rio de Janeiro: RiMA Editora, 2004. v. 1. 376p.

GUARINO, N. Understanding, building and using ontologies. International Journal of Human-Computer Studies, v. 46, p. 293–310, 1997. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1071581996900919>.

GRAHAM, G.; JEREMY, C. 2004. “Resource Description Framework (RDF): Concepts and Abstract Syntax”. W3C. Disponível em: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

MOURA, A.; PORTO, F.; POLTOSI, M. Integrating Ecological Data Using Linked Data Principles. ONTOBRAS-MOST 2012: 156-167, 2012.

NOY, N.; MCGUINNESS, D. Ontology development 101: A guide to creating your first ontology. Development, v. 32, p. 1–25, 2001. Disponível em: <http://protege.stanford.edu/publications/ontology_development/ontology101.pdf>.