Jun 19th, 9:00 AM - 10:20 AM

An Application of geographical and Statistical Linked Data to Ecology: The Brazilian Cerrado Ontology Network and Qualitative Reasoning Models

Adriano Souza  
*University of Brasilia, souzaajb@unb.br*

Oscar Corcho  
*Universidad Politecnica de Madrid, ocorcho@fi.upm.es*

Paulo Salles  
*University of Brasilia, psalles@unb.br*

Luis Vilches-Blázquez  
*Universidad Politecnica de Madrid, lmvilches@fi.upm.es*

Follow this and additional works at: [https://scholarsarchive.byu.edu/iemssconference](https://scholarsarchive.byu.edu/iemssconference)

Part of the [Civil Engineering Commons](https://scholarsarchive.byu.edu/civilengineeringcommon), [Data Storage Systems Commons](https://scholarsarchive.byu.edu/datastoragesystemscommon), [Environmental Engineering Commons](https://scholarsarchive.byu.edu/environmentalengineeringcommon), and the [Other Civil and Environmental Engineering Commons](https://scholarsarchive.byu.edu/othercivilandenvironmentalengineeringcommon)

https://scholarsarchive.byu.edu/iemssconference/2014/Stream-A/46

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
An Application of geographical and Statistical Linked Data to Ecology: The Brazilian Cerrado Ontology Network and Qualitative Reasoning Models

Adriano Souza\textsuperscript{a}, Oscar Corcho\textsuperscript{b}, Paulo Salles\textsuperscript{c} and Luis Vilches-Blázquez\textsuperscript{d}

\textsuperscript{a,c}University of Brasilia, Institute of Biological Sciences, Graduate Program in Ecology, Campus Darcy Ribeiro, CEP: 70904-970, Brasilia, DF, Brazil.
\textsuperscript{b,d}Ontology Engineering Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Spain.
\textsuperscript{a} souzaajb@unb.br; \textsuperscript{b} ocorcho@fi.upm.es; \textsuperscript{c} psalles@unb.br; \textsuperscript{d} lmvilches@fi.upm.es

Abstract: This work aims to link data about wood plant communities obtained from scientific studies about the Brazilian Cerrado biome, meteorological and environmental data, geographical information (maps) and Qualitative Reasoning (QR) conceptual models. An ontology network was developed in order to integrate data from different sources. Data Cube, AEMET, GeoSPARQL, and Time ontologies and QR formalism were used as general ontologies. In addition, two domain ontologies - Cerrado Concepts and Wood Plant Dynamics Ontology (Ccon) and Fire Ontology (Fire) - were developed to represent scientific knowledge about vegetation ecology focused on vegetation dynamics under different burning regimens. Datasets provided by Brazilian government agencies and those obtained from scientific literature were transformed into RDF using LOD Refine and visualized in Map4RDF. The interface integrates map visualization using Google Maps API with facet browsing. Given the aggregated information, the application selects scenarios available in the repository and suggests relevant models for the user. Ongoing work investigates possibilities of automatically integrating simulation results. This research has potential to boost both applications of geographical and statistical linked data technologies to new areas, and perspectives for including results obtained from simulation models into ecological research and applications.

Keywords: Semantics; Metadata and Ontologies of Natural Systems; Geolinked data; Qualitative Reasoning Models; Brazilian Cerrado.

1. INTRODUCTION

Ecological research and its application to natural resources conservation and management require researchers to access data about distinct geographic regions, available on different sources, stored using different technologies (Green et al., 2005). In addition, to integrate and analyze the data; to produce predictive and explanatory models; and eventually to develop public policy recommendations is also required (Green et al., 2005). Villa et al. (2009) point out that as ecological and environmental modeling activities are related to knowledge representation and management, ontologies are increasingly being used to support natural system modelling and to enhance its rigor and consistency (Villa et al., 2009).

In the ecology domain, there are few works developed or under development, such as the Environmental Ontology (EnvO) (“Environment Ontology,” n.d.), the Landscape Ontology (Lepczyk et al., 2008); the Ecological Concepts ontology (Williams et al., 2006) and the Extensible Observation Ontology (OBOE) (Madin et al., 2007). However, current available ontologies do not directly cover issues regarding ecological communities. Accordingly, the present work addresses the semantic formalization of concepts from the domain of ecological communities and processes that can lead to changes in their structure and functioning over time that are not well covered by other ontologies.
A large portion of ecological knowledge can be characterized as incomplete, fuzzy, uncertain, sparse, empirical, and non-formalized. Hence it is often expressed in qualitative terms, verbally or diagrammatically (Salles and Bredeweg, 2006a). This way, formalisms developed in the field of qualitative reasoning (QR) (e.g. Forbus, 1984) can be considered natural candidates to form the basis for conceptual modeling tools to be used, for example, in the ecology domain. In QR-based conceptual models, the quantities that describe the dynamic features of a system hold qualitative information (Bredeweg et al., 2013). They also aid in proposing hypotheses and sacrifice details of the system while emphasizing the general characteristics, and they can show the consequences of what we believe to be true, and can be simulated using computer processing (Nuttle et al., 2009).

In DynaLearn (www.DynaLearn.eu), models can be exported to OWL, and using semantic technology, be further used to provide recommendations based on a repository of models built by experts for students learning how to build models (Bredeweg et al., 2013). Both Garp3 and DynaLearn workbenches have also been extensively used for conceptual modelling in ecology research ((Goulart et al., 2013; Kansou et al., 2013; Salles and Bredeweg, 2006b).

The work described here aims to develop a prototype system based on geographical and statistical linked data initiatives (Vilches-Blázquez et al., 2010a), to be designed to interconnect data from different sources with qualitative reasoning models.

The paper is organized as follows: we start describing the study area, which is followed by the overview of the ontological model used; then a brief description of ontologies reused and developed is presented. After that, we describe how the data was transformed, linked and published. Finally, we discuss related works and the conclusions of the work.

2. THE USE CASE STUDY AREA

The Cerrado is the second largest biome in Brazil with about 2 million km² in the central area of the country. This biome is considered a hotspot of biodiversity (Myers et al., 2000), and about 48% of its original cover was already deforested.

The Cerrado biome presents a remarkable physiognomic variation, spanning from forest like (Cerradão) to open areas (Campo sujo and Campo limpo) with intermediate savanna like physiognomies (Cerrado sensu stricto and Campo cerrado) (Oliveira-Filho and Ratter, 2002). Several studies have been performed aiming to understand the factors governing the past and present geographical distribution of the physiognomies, its structure and composition (Henriques, 2005; Oliveira-Filho and Ratter, 2002). According to Henriques (2005) to understand the dynamics of the vegetation to be able to develop management strategies it is essential to identify the specific mechanism that enables the coexistence and define the density of trees in Cerrado. To approach this problem the present work proposes that semantic technologies and qualitative reasoning modeling can be applied to integrate resources about ecological systems and help ecologists to theoretically explore the hypotheses, such as the demographic bottleneck hypothesis, about mechanisms of tree-grass coexistence in Cerrado biome.

3. ONTOLOGICAL MODEL OVERVIEW

We developed an Ontology Network (Figure 1) to represent the scientific knowledge about vegetation ecology focusing on the dynamics of wood plants of the Brazilian Cerrado Biome. It consists of ontologies from different domains, such as statistics, climate, geospatial, time, qualitative reasoning, plant communities and fire. In this model, we reused existing ontologies that match the knowledge representation requirements acquired over the development of the use case prototype, which are:

- To represent multi-dimensional statistical data and the information observed. This requirements is covered by the RDF Data Cube vocabulary the W3C (World Wide Web Consortium) RDF standard to publish statistical information on the web (Cyganiak and Reynolds, 2013).
- To represent meteorological measurements and their properties. The AEMET network of ontologies (Poveda, 2011) covers this requirement. It is a modular ontology related to meteorological measurements modeling. Each measurement represents the atmospheric condition (humidity, temperature, precipitation etc) at a place and time (Poveda, 2011).
- To represent geographical information and geospatial data. The GeoSPARQL ontology which provides a vocabulary to express geospatial information (Perry and Herring, 2012) covers this requirement. GeoSPARQL defines a vocabulary for representing geospatial data in RDF, and it
defines an extension to the SPARQL query language for processing geospatial data (Perry and Herring, 2012).

- **To represent time, instants and intervals.** This requirement is covered by Time Ontology. Which provides a vocabulary for expressing facts about topological relations among instants and intervals, together with information about durations, and about datetime information (Hobbs and Pan, 2006).

- **To represent qualitative reasoning models.** This requirement is covered by QR formalism ontology defines each of the terms in the Garp3 QR formalism which is used in DynaLearn ILE workbench. The ontology provides the URIs that identify the types of ingredients and is also meant to adequately express the meaning of the QR terms in OWL (Liem, 2013).

- **To represent ecological concepts such as dynamics of plant communities, and the different physiognomies of Cerrado biome.** This requirement is covered by the Cerrado Concepts and Wood Plant Dynamics Ontology (Ccon). It was developed to represent the scientific knowledge about vegetation ecology focused to describe the dynamics of wood plants in the Brazilian Cerrado Domain. The ontology defines the different characteristics of the physiognomies in terms of tree-grass cover, also defines the concept of savanna. It has distinct classes to represent physiognomies. In addition, the ontology includes classes to represent the biological diversity, parameters used to describe the structure and dynamics of wood plant communities and the main determinant factors of the structure and composition of these plant communities.

- **Finally, to represent fire regimes and fire events.** The Fire Ontology (Fire) covers these requirements. The ontology of Fire was created in order to represent the set of concepts about the fire occurring in natural vegetation, its characteristics, causes and effects, with focus on Cerrado vegetation domain. The fire plays a determinant role on the structure and composition of Cerrado physiognomies (Moreira, 1992). Currently, the main cause of burn in Cerrado is anthropogenic, however, the natural lightnings can produce fire in the beginning of the raining season and some evidence suggest the presence of fire in savannas’ history before human influence. Some important effects of fires, such as plant mortality and topkill were considered, mainly those directly related to wood plant dynamics in Cerrado.

![Figure 1](image)

**Figure 1.** Overview of the ontological model.

In the domain ontologies we were able to reuse several terms from Crop-Wild Relatives ontology (CWR) and Environmental Ontology (EnvO). Nevertheless we added specified elements that are typical to wood plant community dynamics, such as *mortality* and *recruitment rates*, and the relationships between these elements and environmental variables, e.g. *affects* and *isCausedBy*. We also specified characteristics of fire in vegetation such as *FireRisk* and *FireFrequency*, and *hasRisk* as an object property. Such elements and properties are not specified in other ontologies in the ecological domain.

We carried out an evaluation of the domain ontologies with experts in ecology and computer science. We prepared two google forms with open and Likert scale questions asking about conceptual correctness of the ontological models, then the forms and the full documentation of the ontologies were sent by e-mail to evaluators, whose answers were collected and analyzed.

---

1 http://cerrado.linkeddata.es/ecology/ccon
2 http://cerrado.linkeddata.es/ecology/fire
According to the experts answers, in spite of further improvements can be implemented, the concepts used in both ontologies were considered, in general, appropriate. They suggested a few changes, such as to add more physiognomy types, to add types of diversity (i.e. structural, functional and compositional diversity). Other comments suggested to change how some classes and subclasses are organized, for example the term Season should not be a subclass of Environment.

4. DATASET IDENTIFICATION AND TRANSFORMATION

We collected open data (Table 1) from Brazilian government agencies such as INMET\(^3\) (Brazilian National Institute of Meteorology), INPE\(^4\) (Brazilian National Institute of Space Research) and IBGE\(^5\) (Brazilian Institute of Geography and Statistics) and LAPIG – UFG\(^6\) (Image Processing and Geoprocessing Lab of Federal University of Goiás). Data about wood plant species occurrence and abundance, from different studies published in scientific journals or theses were also collected. The datasets were found in different formats, such as spreadsheets, text files, shape files. And to perform the transformations of the datasets into RDF we used the LOD Refine. We also used geometry2RDF to transform shape files into RDF.

We built the QR models using DynaLearn software in order to represent the dynamics of Cerrado vegetation. The models were exported to OWL format.

Table 1. The datasets used, their original format and the tools used to transform them into RDF format.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Provenance</th>
<th>File format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vegetation Datasets</td>
<td>Scientific studies</td>
<td>Spreadsheet, PDF</td>
</tr>
<tr>
<td>QR models</td>
<td>-</td>
<td>OWL</td>
</tr>
<tr>
<td>Meteorological data</td>
<td>INMET</td>
<td>Text file</td>
</tr>
<tr>
<td>Fire Occurrence</td>
<td>INPE</td>
<td>Text file</td>
</tr>
<tr>
<td>Brazilian Soils</td>
<td>IBGE</td>
<td>Shape file</td>
</tr>
<tr>
<td>Brazilian Vegetation</td>
<td>IBGE</td>
<td>Shape file</td>
</tr>
<tr>
<td>Brazilian Biomes</td>
<td>IBGE</td>
<td>Shape file</td>
</tr>
<tr>
<td>Remaining areas of Cerrado biome</td>
<td>IBGE</td>
<td>Shape file</td>
</tr>
<tr>
<td>Deforestation scenarios of Cerrado biome</td>
<td>LAPIG - UFG</td>
<td>Shape file</td>
</tr>
</tbody>
</table>

5. URI DESIGN

The base URI we adopted is http://cerrado.linkeddata.es/, this segment is in common to all elements in the knowledge base. TBox, that is, classes and properties, and ABox, instances and components, were separated into http://cerrado.linkeddata.es/ecology/ and http://cerrado.linkeddata.es/resource/ URI schemes, respectively.

The URI scheme of TBox components were also separated into http://cerrado.linkeddata.es/ecology/ccon/ to specify elements that belong to Ccon and http://cerrado.linkeddata.es/ecology/fire/ to the Fire ontology.

6. LINKING AND PUBLISHING DATASETS

We used the Silk Link Discovery Framework\(^7\) to identify and create owl:sameAs links of entities between datasets, models and maps. After that, the sameAs Link Validator tool\(^8\) was used in order to validate the relationships discovered when they apply to geographical data. Using this method, we enrich reference information (geometry) with data.

Virtuoso\(^9\) was used to publish RDF data on the web. It provides a combination of the functionalities of traditional DBMS, virtual databases, RDF triple stores, XML stores, web application

---

\(^3\) http://www.inmet.gov.br/
\(^4\) http://www.inpe.br/
\(^5\) http://www.ibge.gov.br/
\(^6\) http://www.lapig.iesa.ufg.br/ lapig/
\(^7\) http://www4.wiwiiss.fu-berlin.de/bizer/silk/
\(^8\) http://oeg-dev.dia.fi.upm.es:8080/sameAs/
\(^9\) http://virtuoso.openlinksw.com/
servers and file servers (Vilches-Blázquez et al., 2010b). Pubby\textsuperscript{10} is used to visualize and navigate over the raw RDF data on top of Virtuoso.

7. EXPLOITATION

We deployed a web based application, Map4RDF\textsuperscript{11}, to provide a visualization of the aggregated information. Its interface integrates map visualization using Google Maps API with facet browsing. The application renders and displays on the map the distinct geometrical shapes of the features published as RDF. In addition, it is possible to display statistical data on the map so that the user can observe and compare them. Our goal is to connect qualitative reasoning models to environmental data available in the dataset, and run the simulations in order to explain why a specific situation came up or to predict possible outcomes. In the current status of implementation, when the user selects a single variable or a set of variables within the aggregated datasets, the application selects among the models available in the repository those in which the same variable or set of variables is included.

The algorithm basically uses the “owl:sameAs” links generated between the variable in the dataset and the quantity in one or more models. If at least one of such links is established, then the model will be presented as a candidate to the user.

To illustrate this mechanism we built six qualitative models in DynaLearn workbench. All of them present the Entities ‘Wood Layer’ and ‘Ground Layer’, and both entities are associated to the quantities Mortality rate, Recruitment rate, Number of individuals and Density. Besides these four quantities, the models may include one or more environmental factors (temperature, relative humidity, precipitation, fire risk and underground water), as shown in Table 2. The complete model contains these five variables affecting the mortality and recruitment rates of plants in the Wood Layer and Ground Layer.

<table>
<thead>
<tr>
<th>Model name</th>
<th>Quantities included* (Environmental variables)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Veg dyn only precipitation</td>
<td>Precipitation</td>
</tr>
<tr>
<td>Veg dyn only temperature</td>
<td>Temperature</td>
</tr>
<tr>
<td>Veg dyn only relative humidity</td>
<td>Relative humidity</td>
</tr>
<tr>
<td>Veg dyn only fire risk</td>
<td>Fire risk</td>
</tr>
<tr>
<td>Veg dyn complete</td>
<td>Temperature; Relative humidity; Precipitation; Fire risk</td>
</tr>
<tr>
<td>Veg dyn underground water and fire risk</td>
<td>Underground water; Fire risk</td>
</tr>
</tbody>
</table>

* All the models include quantities related to the vegetation (species, number of individuals, mortality etc.)

After having all the models and datasets (Meteorological data, Fire occurrence data, and Vegetation dynamics data) stored in the server we perform the generation of links between the datasets and models. We use the SPARQL endpoint to run the construct query and establish owl:sameAs links. Thus using this query the system associates variables present in the datasets to variables present in the models. For example, the user wants to inspect the variable ‘Precipitation’, which is present in the Meteorological dataset and it is further linked only with the quantity Precipitation present in the models “Veg dyn only precipitation” (Figure 2) and “Veg dyn complete”, and both models are presented to the user. Although the variables ‘Temperature’ and ‘Relative humidity’ were not mentioned by the user, these two quantities are included in the model “Veg dyn complete”.

Now suppose the user wants to analyze the effects of the variable ‘Fire risk’, present in the Fire occurrence datasets, which is linked with the quantity Fire risk found in the models “Veg dyn only fire risk” and “Veg dyn underground water and fire risk” (Figure 3). In this case, the situation is different, as there is no data available in the datasets to provide values to the quantity Underground water, found in the model “Veg dyn underground water and fire risk”. However, this quantity has assigned the qualitative value <critic, increasing>, as shown in Figure 3. This way, it is possible to run simulations with the model, as all the quantities have initial values defined either by numerical databases or qualitative values assigned by the application or the user.

\textsuperscript{10} http://www4.wiwiss.fu-berlin.de/pubby
\textsuperscript{11} http://oegdev.dia.fi.upm.es/map4rdf/
Figure 2. Illustration of how the "owl:sameAs" links generated between datasets and a model, in this case the variable Number of individuals is found in the "Vegetation dataset" and as a quantity in the model. In addition the variable Precipitation found in the "Meteorological dataset" is also found as a quantity in the model. On the other hand, this model has no owl:sameAs link with "Fire occurrence dataset", because of the lack of common variables between them.

Figure 3. Links "owl:sameAs" generated between model "Veg dyn underground water and fire risk" and datasets. No links with meteorological data was set and the variable Underground water has no correspondence with any dataset. However, qualitative values (<critic, increasing>) are already assigned to this quantity, so the model is complete for simulations.

8. DISCUSSION

This study is based on the work performed by GeoLinkedData.es, publishing diverse information sources belonging to the National Geographic Institute of Spain (Vilches-Blázquez et al., 2013). The capability to interoperate, browse and explore the data inspired this work to use ecological heterogeneous data of Brazilian Cerrado plant communities in a geospatial scale and, in addition, as shown in the present work, to link these data to qualitative ecological models using semantic techniques. This way the system is able to select an appropriate model based on the set of data selected or filtered by the user.

The implementation of the facility for selecting qualitative models follows four principles: (a) if there is at least one common variable between the aggregated dataset selected by the user and a model in the repository, the link "owl:sameAs" is generated and the model is selected and presented to the user as a candidate for simulation; (b) if the user selects a variable in the aggregated datasets, and this variable is a quantity of a model, which is then selected, and besides that, this model includes one or more quantities that were not filtered by the user but are also found in the datasets, then to these extra quantities have also assigned numerical values so they can be used for the simulations; (c) if a situation similar to the one described in the previous item (b) occurs, but the extra
quantities are not found in the datasets, the simulation may be possible if qualitative values are assigned to these extra quantities, and (d) if the user selects a variable from the datasets that does not exist as a quantity in any model, then it does not select any model from the repository.

The current version of the application is able to follow these four principles, and selects one or more models ready for simulation. However, the simulation run and the interpretation of the results have to be done manually. The results reported here are seem as a first step in the association between datasets and qualitative models.

Ongoing work investigates means to select and run simulations automatically with the application. In order to do so, it is necessary to map numerical data into relevant points and intervals, establishing a correspondence between numerical data and qualitative values. Besides that, it is necessary to automatically produce an interpretation of the meaning of direct influences and qualitative proportionalities, so that the causal model could be read and the simulation results interpreted for the user.

In the present study we describe the use of semantic techniques to create connections between datasets and qualitative reasoning models. Considering that providing conceptual frameworks to express the realm of natural systems knowledge is far from trivial, and natural systems sciences need to address the most complex domains and are still far from the general acceptance (Villa et al. 2009), the present work is a contribution to the progressive adoption of such ontology-driven approaches.

9. CONCLUSIONS

In this study we explore a linked data approach to link qualitative reasoning models to wood plant species occurrence and dynamics, fire occurrence, Brazilian geographical information, and meteorological data, provided by Brazilian governmental agencies and scientific studies. We followed a methodology based in the Geolinked data approach aiming to investigate the application of linked data principles to ecology.

This work contributes to:
(1) Formalize into ontologies knowledge about Cerrado biome, wood plant species dynamics and fire regimes;
(2) Semantically enrich datasets about wood plants of Cerrado, fire occurrence, meteorological data, and maps of soils, vegetation, climate, and deforestation scenarios.
(3) Create links between data of different studies from distinct locations adding geospatial information to wood plant species occurrence and dynamics;
(4) Enable an alternative method based on semantic technologies to search for biodiversity patterns;
(5) Enable publish and link biodiversity data not only in terms of species occurrence, but also in terms of number of individuals over different periods of time, that is variation in the population sizes according to environmental variables or conditions;
(6) Enable to approach questions with regard to tree cover in Cerrado biome using a visual tool including geospatial dimension;
(7) Create links among species occurrence, environmental variables such as soil types, temperature, precipitation, fire events and deforestation, and qualitative reasoning models, that can used for simulations and as such to support predictions and explanations about Cerrado behavior.

The topics addressed in this research have potential to boost both applications of geolinked data technologies to new areas, and to open new perspectives for research involving ecological data management, integration and use.

ACKNOWLEDGEMENTS

We would like to acknowledge CAPES-PDSE (Programa de Doutorado Sanduíche no Exterior) for the doctoral fellowship granted to the first author, for the support of the Ontology Engineering Group (OEG) members at UPM, and the DynaLearn Project (co-funded by the EC within FP7, Project no. 231526).

REFERENCES


Hobbs, J.R., Pan, F., 2006. Time Ontology in OWL. URL http://www.w3.org/TR/owl-time/


