Implementing an Architecture for Semantic Search Systems for Retrieving Information in Biodiversity Repositories

Flor Karina Mamani Amanqui[1], Kleberson Junio Serique[1], Franco Lamping[1], José Laurindo Campos dos Santos[2], Andréa Corrêa Flôres Albuquerque[2], Dilvan de Abreu Moreira[1]

[1] University of São Paulo, Computer Science Dept, São Carlos, Brazil
{flork, serique, lamping, dilvan}@icmc.usp.br
[2] National Institute for Amazonian Research, Manaus, Brazil
{andreaa, lcampos}@inpa.gov.br

Abstract. Biological diversity is of essential value to life sustainability on Earth and motivates many efforts to collect data about species, giving rise to a large amount of information. Biodiversity data, in most cases, is stored in relational databases. Researchers use this data to extract knowledge and share their new discoveries about living things. However, nowadays the traditional search approach, based on keywords, is not appropriate to be used in large amounts of heterogeneous biodiversity data. In addition, the search by keyword has low precision and recall in this kind of data. In this paper, we present a novel architecture, for ontology based semantic search systems, and test results of a prototype system, implemented using state of the art free semantic web tools, using a set of representative data about biodiversity from INPA (consisting of specimens of fish and insects). This test results show that the prototype had better recall and precision than keyword based methods (for the same dataset). In the semantic web, ontologies allow knowledge to be organised into conceptual spaces in accordance to its meaning. For that reason, for semantic search to work, a key point is to create mappings between the data, stored in relational databases, and the ontologies describing this data. This work also developed such a mapping.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords: Biodiversity, Ontology, Data Integration, Semantic Search

1. INTRODUCTION

The Web became one of the most common ways to get scientific data for the acquisition of new knowledge. However, this wide data availability generates large amounts of information in all areas of knowledge. Due to the huge size of the Web, it has become difficult for users to find the information they need.

In the biodiversity field, it is not different. There is a large quantity of online data generated by research institutions that have collections of specimens collected in Brazil. Efforts, like the SpeciesLink network or the Global Biodiversity Information Facility (GBIF)[1], integrate specie and specimen data available in natural history museums, herbaria and culture collections, making it openly and freely available on the Internet. They integrate organisations like the National Institute for Amazonian Researsh (INPA)[2], Pará's Emílio Goeldi Museum (MPEG)[3], the Amazon Biotechnology Center (CBA)[4],

[1] http://www.gbif.org/
[2] INPA http://www.inpa.gov.br
[3] http://www.museu-goeldi.br
[4] http://www.suframa.gov.br/cba/

the Reference Center for Environmental Information (CRIA)[5], and The New York Botanical Garden (NYGB)[6], among other important organisations for dissemination of biological data collections. That has led to discussions about the best way to arrange this data and provide tools and environments that stimulate and facilitate information sharing.

This proliferation of information from different sources means that the search for information could be met by a variety of available resources, which store data about the same domains but have different characteristics. Therefore, the need for integration and analysis of data from these sources becomes more evident.

For biodiversity data to be analysed and retrieved efficiently, it is necessary to use new technologies to improve human and computer collaboration on the Web. Interestingly, despite improvements in search engine technology, the difficulties remain essentially the same. It seems that the amount of Web content outpaces technological progress. Fortunately, Semantic Web technologies enable explicit declaration of knowledge embedded in many Web-based applications, integrating information in an intelligent way, providing semantic-based access to the Web.

Berners-Lee, Hendler and Lassila introduced the concept of the Semantic Web in 2001 and helped to stimulate innovative ideas and new technologies from different computing areas [Berners-Lee et al. 2001]. There are a number of important issues related to the Semantic Web: ontologies, languages for the Semantic Web, semantic search, semantic markup of pages and services.

In this work, we present an architecture for a Semantic Search system that supports mapping between the data (INPA's biodiversity data) stored in relational databases, and the biodiversity ontology (OntoBio). OntoBio was developed by INPA. Its main objective is to provide a clear and precise conceptualisation of the information presented in biodiversity data collections. The complete ontology is presented in details in [Albuquerque 2011].

Futhermore, the mapping was implemented using free state of the art Semantic Web tools and tested on a set of representative data about biodiversity from INPA. The contributions of this work are as follows: (i)A proposal for a new Semantic Search Architecture for biodiversity data; (ii)A mapping between the INPA's biodiversity data (stored in relational databases), and the ontology OntoBio from INPA and (iii) finally, experimental results on collected data from INPA showing improvements in precision and recall when compared to keyword based methods.

The remainder of this article proceeds as follows: Section 2 presents our Semantic Search Architecture. Section 3 presents the experiments and evaluations. Section 4 describes related works, and in Section 5 the conclusions, discussions and future works are presented.

2. RELATED WORK

This section focuses on the analysis of search mechanisms for integrating information about biodiversity. In Brazil, there are three important web tools that manage biological collections: SpeciesLink[7], Specify[8] and SinBiota[9]. In our approach, we have analyzed these systems to understand the needs and requirements of biodiversity experts. SpeciesLink is used for the comparison between keyword based search with semantic based search because it stores information from INPA collections.

In the context of semantic web technologies applied to biodiversity data, there is a large number of projects implemented and published that supports research on biodiversity conservation on the Web. A number of techniques have been developed for use ontologies to retrieve relevant documents

[5] http://www.cria.org.br/
[6] http://www.nybg.org
[7] http://splink.cria.org.br/
[8] http://informatics.biodiversity.ku.edu/
[9] http://sinbiota.cria.org.br/

in response to a query as [Kara et al. 2012], [Freitas et al. 2012], [de S. Fedel et al. 2012], [dos Santos et al. 2011]. These systems use relational databases to store the biodiversity data and ontologies. However, none of the work focused on the problem of storage and retrieval of Resource Description Framework (RDF)[10] triples. Also, an additional limitation in many of the existing approaches is the lack of an an evaluation of the quality of results.

3. ARCHITECTURE FOR SEMANTIC SEARCH

The Semantic Search Architecture for the biodiversity domain created by us is shown in Figure 1 and its components are detailed next.

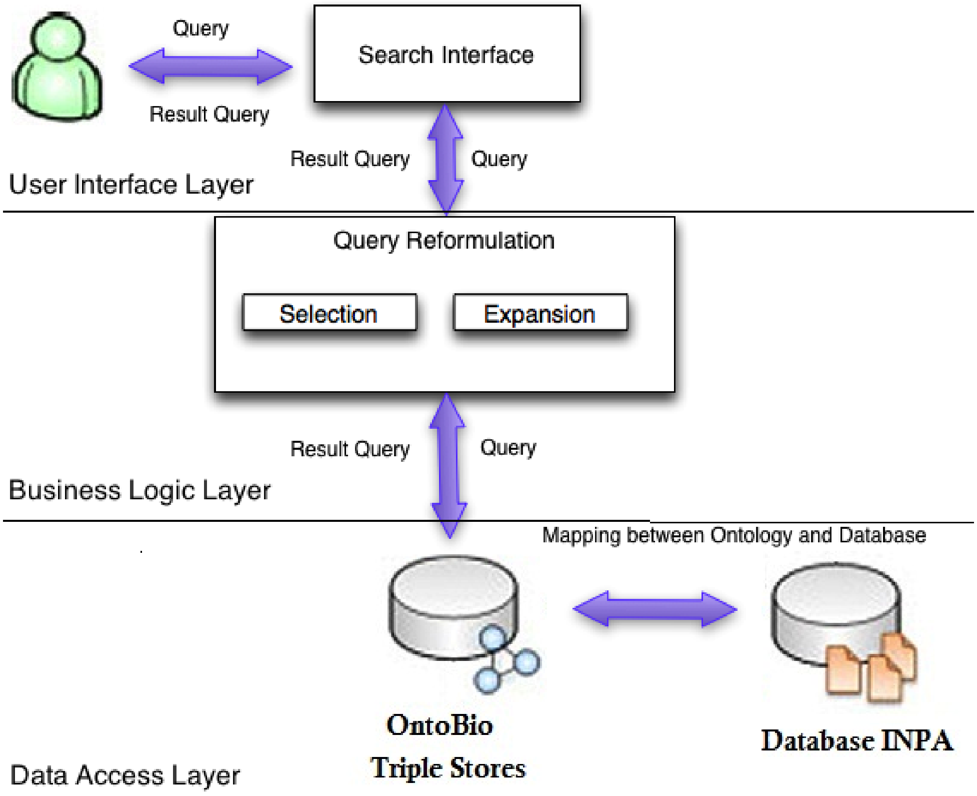


Fig. 1. Architecture for Semantic Search

The User Interface layer is responsible for the interaction between user and system. The search process begins with an initial keyword list, entered by the biodiversity specialist, which represents his/her search intentions.

The Business logic layer processes the information retrieval. Its components are detailed as follow.

The Query Reformulation Component receives the input of search terms from the user, selects and expands keyword lists by adding semantically related terms, using the techniques of expansion

[10] http://www.w3.org/RDF/

and semantic similarity. Querying the semantic data is simplified because of the use of triples. These triples are generated when the database records are mapped into ontology concepts.

The Data Access layer provides access to the RDF triples stored in the Triple Store (in our prototype, Virtuoso) for the layers above and also to machine clients in the web through a SPARQL Endpoint.

The Mapping Component The mapping of the database records into RDF triples is done offline. It generates the triples that will be stored in the triple store (in our case, Virtuoso) and queried during user searches. The Mapping Component loads the domain ontologies, taxonomic information and the collection database and transforms them in a set of RDF triples. We used Ontop, a platform to query databases as Virtual RDF Graphs using SPARQL, to do the mapping between the relational databases records and the OWL ontologies. Ontop has two tools: (i)**OntopPro**, which is a Protege 4 plugin that implements a graphic mapping editor to allow the definition of data sources (databases) for ontologies and the mapping of records, from this data sources, into ontology entities using an intuitive mapping language [Mariano R and Calvanese 2012]. (ii)**Quest**, which is a SPARQL query engine/reasoner that supports RDFS and OWL 2 QL entailment regimes and SPARQL-to-SQL query rewriting. Using OntopPro, it is possible to export the RDF triples generated by the Quest tool to a file.

3.1 Semantic Search Algorithm

The basic idea of our algorithm is to compare input keyword with OntoBio resources (subject, predicate and object) in Virtuoso triple store. Virtuoso supports the SPARQL query language and provides a faceted browser user interface for querying the RDF store. The use of triple stores makes our semantic search architecture more versatile due to the fact that they allow great volumes of data and work directly with the SPARQL language. The algorithm used is shown below:

```
Algorithm: SemanticSearch (String Query)
Step 1: Establish connection with Virtuoso Triple Store and
OntoBio ontology
Step 2: Extract OntoBio information (hierarchy)
Step 3: For each Query in OntoBio
(a) Construct SPARQL query with Query as object, subject or
predicate
(b) Calculate similarity of Query with subject, object and
predicate from OntoBio
Step 4: Go to Step 3 until no string is left in Triples OntoBio
or no more text are to be considered
Step 5: Arrangue the results according to order of similarity
Step 6: Send and display the results
```

4. EXPERIMENTS

We implement a prototype using Java, Google Web Toolkit 2.5.1 (to create a web application), Jena RDF framework (to process SPARQL queries) and Virtuoso Server, as triple store. In order to validate our architecture, researchers from our group and biodiversity scientists (from INPA) were interviewed to categorise important information from the INPA dataset (e.g. genus, family, specie, location description). We then defined use cases with features and scenarios to identify important user tasks. These use cases are:

USE CASE 01: Statistics relating to fishing and use of resources

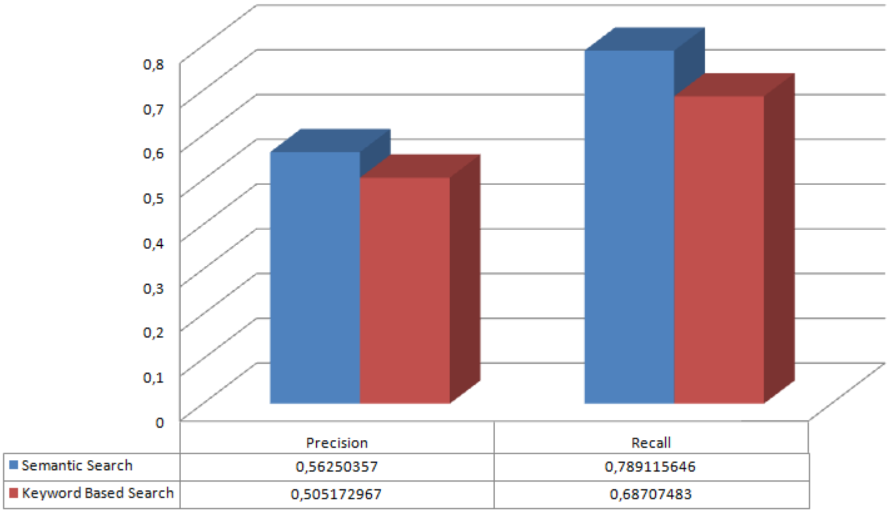
INITIATOR: Mary Smith, Biologist, 29 years-old **GOAL:** To determine the best areas for aquaculture development, considering different types of species and time of year. Aquaculture is an activity ecologically sustainable that generates income and can propel the economy of a region. **BEST SCENARIO** Mary is working in her laboratory and needs to search fish species suitable for aquaculture in the Amazon. She needs to specify the name of the fish species she is interested and determine the fish size, viable habitats, municipalities (where these habitats are present).

USE CASE 02: Environmental impact study

INITIATOR: John Rubin, Environmental Engineer, 37 years. **GOAL:** To determine if specimens sampled in a swamp are endemic or cosmopolitan. **BEST SCENARIO** To discover, in all collections available, the records of specimens belonging to the same species of the ones found in the swamp; find out the geographical location and the habitat of each collected specimen and then determine if all species found in the swamp are found in different habitats and/or distributed geographic locations.

Using to the previous use cases, we identified the user requirements (for each case) and created samples of the kind of queries they would do. Using just fish and insect data we tested this queries using keyword and semantic search. Experts in biodiversity identified, in each query, what they judged as relevant and non relevant (precision) and if the query returned all relevant data (recall).

We compared the result returned by two search system: our semantic search prototype and the keyword based search from the SpeciesLink [11] tool with data from INPA. The efficiency of the result for each approach was evaluated using the average precision and recall. This evaluation is shown in Figure 2. We used a total of 14 queries (7 for each system). We would like to have done more queries, but the analysis of each query represents a lot of work for the biodiversity experts from INPA, they have to determine, from a big dataset, which records are relevant or not for each query to analyse the quality of the query results. Unfortunately, the time these experts had to dedicate to our project was very limited.



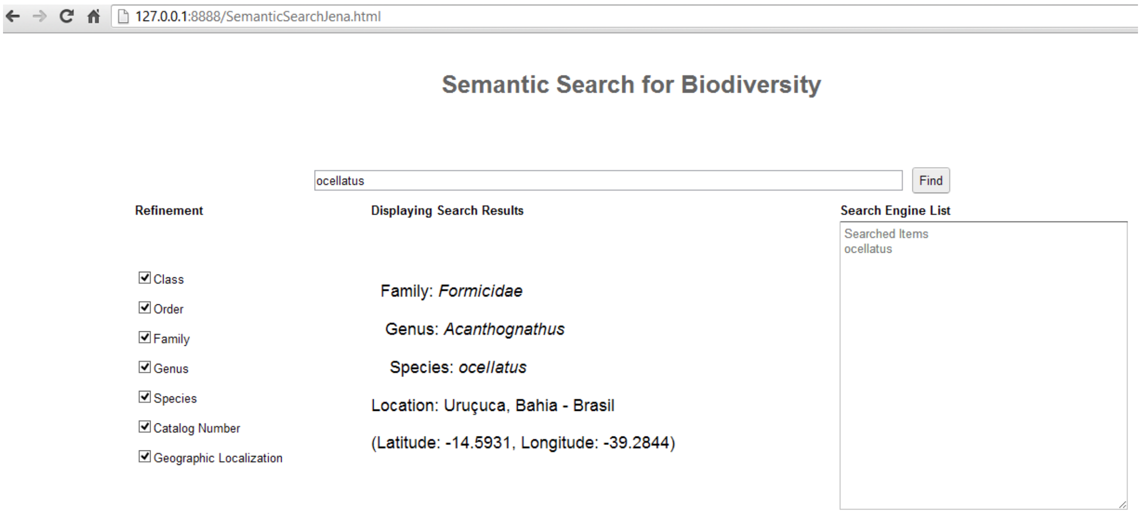
	Precision	Recall
🟦 Semantic Search	0,56250357	0,789115646
🟥 Keyword Based Search	0,505172967	0,68707483

Fig. 2. Evaluation of semantic search with keyword based search

The semantic search architecture had the highest precision (on average), 11.3 % better than using the keyword search. The semantic search recall (on average) was 14.8% better. These results show the advantage of the semantic approach. It is important to note that this tests do not cover cases where

[11] http://www.splink.org.br/

queries need information from different data sources, something almost impossible to do automatically with web based keyword search systems and very easily done using semantic search in SPARQL Endpoints. Figure 3 shows graphic interface to support the user queries.



Semantic Search for Biodiversity

ocellatus

Refinement

☑ Class
☑ Order
☑ Family
☑ Genus
☑ Species
☑ Catalog Number
☑ Geographic Localization

Displaying Search Results

Family: *Formicidae*

Genus: *Acanthognathus*

Species: *ocellatus*

Location: Uruçuca, Bahia - Brasil

(Latitude: -14.5931, Longitude: -39.2844)

Search Engine List

Searched Items
ocellatus

Fig. 3. Screen copy of the prototype using mapping implementations

5. CONCLUSIONS AND FUTURE WORKS

In this article we presented a semantic search architecture that provides a new document retrieval process by exploring a domain knowledge (OntoBio) and semantic search to support scientists in the process of discovery and integration biodiversity data. Our goals were to specify and developed semantic search approaches that allow biodiversity researchers to easily find and access relevant data sources. These approaches can improve precision and recall compared to the existing keyword-based queries offered by tools such as SpeciesLink using the same data from INPA. As future work, we plan to refine the results from our use cases. We intend to reuse geoSPARQL ontology terms to describe georreferenced data. We also intend to extend our current implementation with more advanced structured searches in collaboration with researches from INPA.

The authors would like to thank CNPq for funding this work.

REFERENCES

Albuquerque, A. *Desenvolvimento de uma Ontologia de Domínio para Modelagem de Biodiversidade.* Dissertação de Mestrado. Universidade Federal do Amazonas, 2011.

Berners-Lee, T., Hendler, J., and Lassila, O. The semantic web. *Scientific American* 284 (5): 34–43, May, 2001.

de S. Fedel, G., Medeiros, C. B., and dos Santos, J. A. Sinimbu- multimodal queries to support biodiversity studies. In *Proceedings of the 12th international conference on Computational Science and Its Applications - Volume Part I*. ICCSA12. Springer-Verlag, Berlin, Heidelberg, pp. 620–634, 2012.

dos Santos, V., Baiao, F., and Tanaka, A. An architecture to support information sources discovery through semantic search. pp. 276 –282, 2011.

Freitas, A., Curry, E., and O'Riain, S. A Distributional Approach for Terminology-Level Semantic Search on the Linked Data Web. In *27th ACM Symposium On Applied Computing (SAC 2012)*. ACM Press, 2012.

Kara, S., Alan, O., Sabuncu, O., Akpnar, S., Cicekli, N. K., and Alpaslan, F. N. An ontology-based retrieval system using semantic indexing. *Information Systems* 37 (4): 294–305, June, 2012.

Mariano R, M. and Calvanese, D. *Quest, an OWL 2 QL Reasoner for Ontology-Based Data Access.* KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, Bolzano, Italy, 2012.