Semantic Search Architecture for Retrieving Information in Biodiversity Repositories

Flor K. Amanqui[1], Kleberson J. Serique[1], Franco Lamping[1], Andréa C. F. Albuquerque2[2], José L. C. Dos Santos[2], Dilvan A. Moreira[1]

[1]University of São Paulo (USP) – CEP: 13566-590 – São Carlos – SP – Brazil

[2]National Institute for Amazonian Research (INPA)
CEP.: 69060-001 – Manaus – AM – Brazil

{flork,serique,dilvan}@icmc.usp.br, lamping@grad.icmc.usp.br

andreaalb.1993@gmail.com, lcampos@inpa.gov.br

Abstract. *The amount of biological data available electronically is increasing at a rapid rate; for instance, over 16.500 specimens are available today in the National Institute for Amazonian Research (INPA) collections. However, this data is not semantically categorized and stored and thus is difficult to search. To tackle this problem, we present a semantic search architecture, implemented using state of the art semantic web tools, and test it on a set of representative data about biodiversity from INPA. This paper describes how the mechanism of mapping is designed so that the semantic search can find information, based on ontologies. We show a series of SPARQL queries and explain how the mapping mechanism works. Our experiments, using a prototype of the proposed architecture, showed that the prototype had better precision and recall then traditional keyword based search engines.*

Keywords: Biodiversity, Ontology, Data Integration, Semantic Search

1. Introduction

Biological diversity, or biodiversity, is the term given to the variety of life on Earth. Biodiversity is the combination of life forms and their interactions with one another, and with the physical environment that has made Earth habitable.

The biodiversity information that can be obtained via Internet continues to grow significantly. Every day, new collections, databases, and applications are being added. This information is stored in a variety of formats (spreadsheets, html, xml, pdf and catalogues, amongst others). 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. For that reason, much of this information is never found. The need for integration and analysis of biodiversity information becomes evident.

In this context, finding relevant and recent information is a hard task that is not particularly well supported by current biodiversity software tools. Keyword-based search have serious problems associated with its use: low or no recall; high recall, low precision; initial keywords in search often do not get the wanted results.

The semantic web (an extension of the current Web) tries to represent information in such a way that it can be used by machines, not just for display purposes, but also for automation, integration and reuse across applications [Boley et al. 2001].

There are a number of important technologies related to the Semantic Web: ontologies, languages for the Semantic Web, semantic search, semantic markup of pages and services (that the Semantic Web is supposed to provide). Ontologies, one of the most important ones, are implemented in the RDF(S) (Resource Description Framework/Schema) and OWL (Web Ontology Languages) languages, two W3C recommended data representation models.

In this article, we propose a semantic search architecture that supports mapping between biodiversity data, from INPA's (National Institute for Amazonian Research) collections, stored in relational databases and the ontologies describing it.

The rest of this article is organized as follows: Section 2 describes related works. Section 3 describes our biodiversity ontology. Section 4 presents our semantic search architecture. Section 5 presents a synopsis of our experiments and Section 6 concludes by summarizing our results and describing future works.

2. Related Works

Researchers have proposed various techniques and approaches designed to perform semantic search. We studied a number of them that could be used in the area of biodiversity.

In [Xiong et al. 2009], a method of search based on a smart query agent is proposed (Geoonto). It retrieves information from data catalogs/databases using ontologies. This method associates semantic information in the search process, and generates a refined query string.

In [Latiri et al. 2012], an automatic method of query expansion is proposed in which user requests are expressed in natural language.

In [Mittal et al. 2010], a method hybrid of personalized web information is proposed in which ontology for retrieval of user context is used and a user profile is being maintained.

In [Li and Yang 2008], a method to construct a semantic search engine is proposed. It provides a uniform platform to search, view and operate spatial on information.

In [Santos et al. 2011], an architecture to support semantic search in a metadata repository is proposed. This work discover similar concepts even when different terms are used in their designation or description, since a domain ontology is used to annotate information sources and to expand the user query with terms from the universe of discourse.

A number of techniques have been developed for using ontologies to retrieve relevant documents in response to a query. However, none of them focused on the problem of storage and retrieval of RDF triples. Most of these techniques require complex analysis, involving natural language processing, to discover the context and semantics of query terms. Also, an additional limitation, in many of the existing approaches, is the lack of a quality evaluation of results.

We have developed a semantic search application that uses key semantic web concepts for information retrieval and also technologies such as mapping, triple store and SPARQL queries.

3. The Biodiversity ontology

OntoBio is a biodiversity ontology developed by INPA and UFAM (Federal University of Amazonas) and extended by USP (University of São Paulo). Its main objective is to provide a clear and precise conceptualization of the aspects considered in biodiversity data collection, regardless of a specific application.

The original version of OntoBio is presented in details at [Albuquerque 2011]. One of the advantages of having data annotated using OntoBio concepts is that it can be reused as Linked Data. Linked Data describes a method of publishing structured data so that it can be interlinked and become more useful [Kauppinen and de Espindola 2011].

To better archive that, data annotated using OntoBio has to be easily interlinked with other biodiversity data, already available on the web (as part of the wider Linked Data community), through the use of as many shared concepts as possible. With that in mind, we rewrote the first version of OntoBio to reuse, whenever possible, terms from other public available ontologies to allow better ”linkability” with data already annotated using them.

We added terms from the following public ontologies:

- The Phenotypic Quality Ontology[PATO 2010], which is an ontology of phenotypic qualities, intended for use in a number of applications, primarily defining composite phenotypes and phenotype annotation;
- Basic Geo Vocabulary [WGS84 2003] is a basic RDF vocabulary that provides the semantic web community with a namespace for representing lat(itude), long(itude) and other information about spatially-located things, using WGS84 as a reference datum, and;
- The Geoname Ontology[GeoNames 2011] makes it possible to add geospatial semantic information to the Word Wide Web. All over 8.3 million geonames toponyms now have a unique URL with a corresponding RDF web service.

The OntoBio ontology is presented in the Figure 1. The Protégé 4 ontology editor was used to write the OntoBio ontology in OWL 2 DL. The new version of OntoBio is available through the NCBO's Bioportal http://bioportal.bioontology.org/ontologies/50517. There, users can download, browse and suggest terms for the ontology.

4. An Architecture for Semantic Search

We have proposed the architecture of a semantic search that follows the mechanism of mapping between OntoBio domain ontology, and Database from INPA the collections of insects, fishes, and mammals.

The system overall architecture is shown in Figure 2. It consists of four basic modules: User Interface Layer, Query Reformulation, Mapping Component and Data Access Layer.

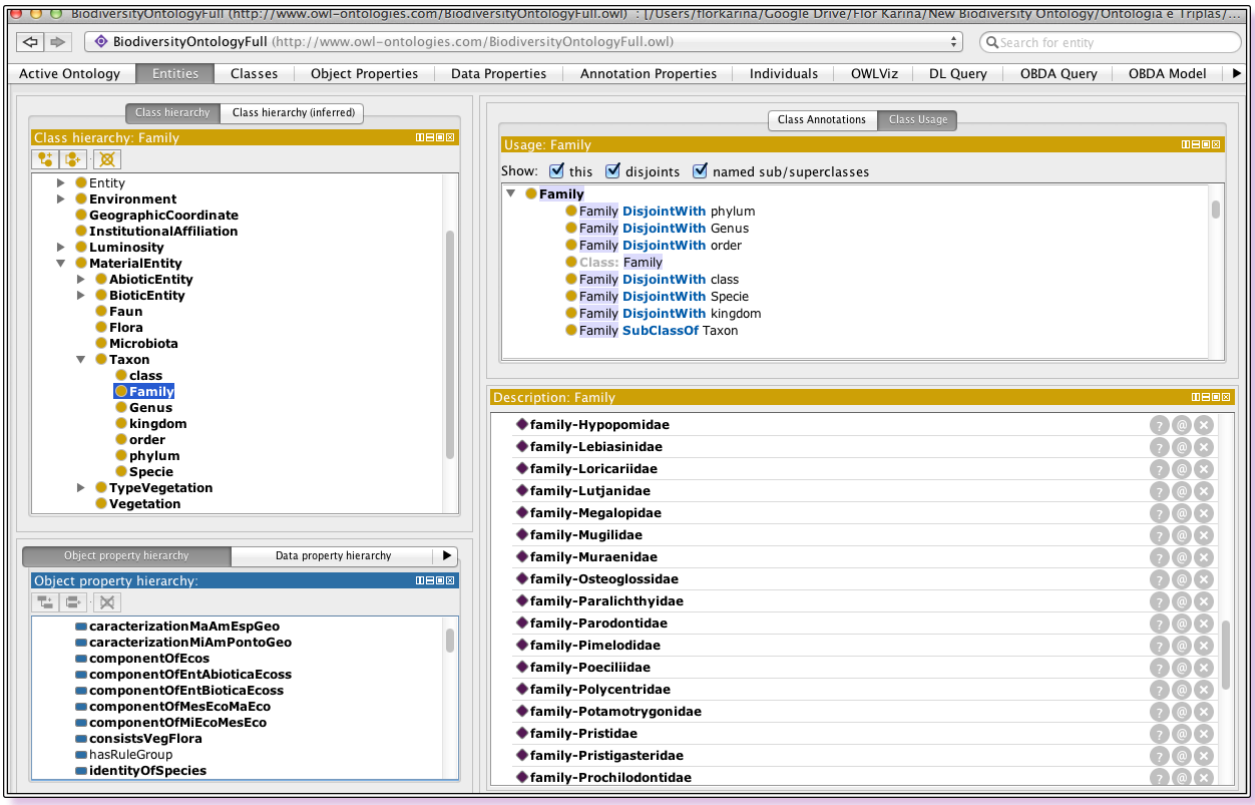

Figure 1. New version of biodiversity ontology

1. **User Interface Layer** is responsible for the interaction between users and system. The search process begins with an initial keyword list, entered by the users, that represents his/her search intentions.
2. **The Query Reformulation component** receives the input of search terms from the user, selects and expands keyword lists by adding semantically related terms, using techniques of expansion and semantic similarity. It uses the SPARQL Writer component to take keyword lists and generate SPARQL queries from them. It uses an algorithm that will be described on the following sections.
3. **The Mapping Component** loads the domain ontologies, taxonomic information and the collection database and transforms them in a set of Resource Description Framework (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 is a platform to query databases as Virtual RDF Graphs using SPARQL. It does the mapping between the relational databases records and the OWL ontologies. Ontop has two tools: OntopPro, which is a Protege 4 plugin that implements a graphic mapping editor; and Quest, which is a SPARQL query engine/reasoner that supports RDFS and OWL 2 QL entailment regimes and SPARQL-to-SQL query rewriting (Mariano R and Calvanese, 2012). The mapping process is divided in three steps:
   (a) **Creation of Mapping Axioms:** OntopPro mappings are done using mapping axioms. A mapping axiom is defined by an SQL query and an ABox

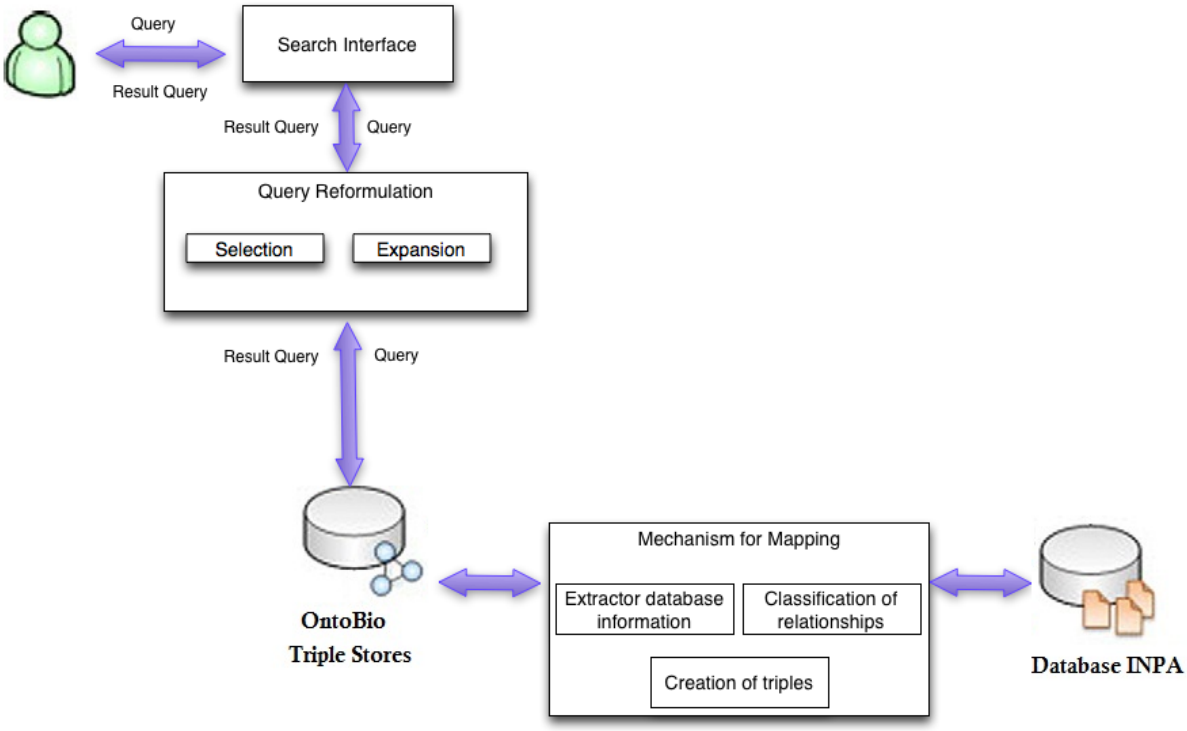

Figure 2. Architecture for Semantic Search

assertion template (Figure 3). An Abox assertion template is a set of RDF/OWL triples, written in a turtle-like syntax, in which the subject and object of the triples allow for variables that reference columns of the SQL query result [Mariano R and Calvanese 2012].
In other words, a mapping axiom defines how the values in each row of the results (of an SQL query) can be used to generate a set of ABox assertions. The mapping axioms were created using information from the OntoBio ontology and INPA experts. Each mapping must contain one or more mapping axioms. Figure 3 shows a valid mapping.

(b) **Generation of RDF Triples:** Mapping axioms generate RDF triples. This generation is done using the Quest tool from Ontop. The Quest reasoner uses query-rewriting techniques to generate triples. The triples are created by replacing the placeholders in the target with the values from the SQL row.

(c) **RDF Triples Loader:** Using OntopPro, it is possible to export the RDF triples generated by the Quest tool to a file. That file is then loaded into the Virtuoso triple store, which is now ready to answer queries using them. The Mapping Component can repeat the process described here, whenever INPA releases updates to its collection records.

4. **Data Access layer** that is the architecture layer that provides access to the RDF triples stored in the Virtuoso Triple Store, using SPARQL, both for the layer above it and for other machines on the network. Triple Store is the common name given to a database management system for RDF Data.

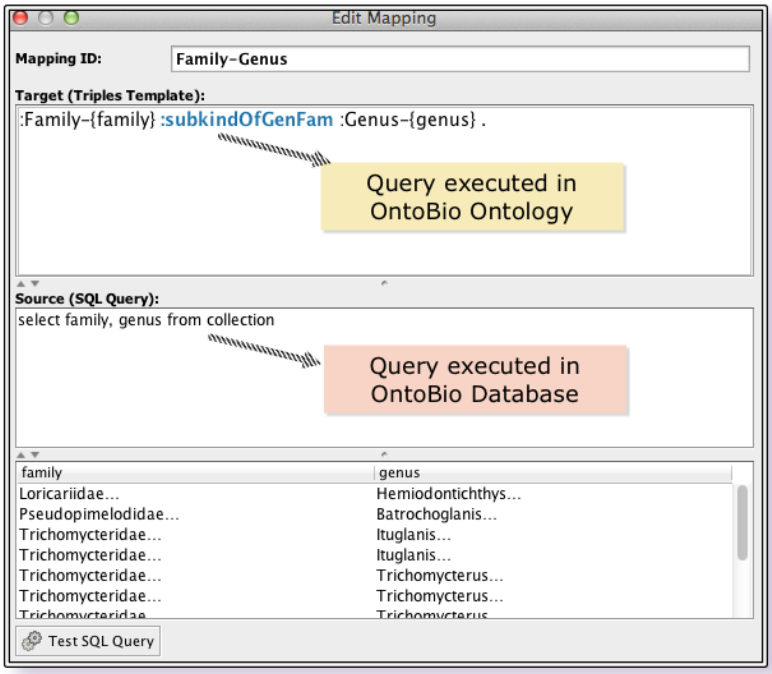

Figure 3. Mapping axiom

4.1. Semantic Search Algorithm

The basic idea of our algorithm is to compare input keywords with OntoBio resources (subject, predicate and object) in the Virtuoso triple store. The Virtuoso platform was chosen because it can store the triples generated from INPA data and work with multiple graphs at the same time.

```
I/P – String nameClass, name of the class selected by user
O/P – String result, result of the queries
BEGIN
Step 1: Establish connection with Virtuoso Triple Store
and ontology OntoBio
Step 2: extracts OntoBio information and find hierarchy
Attribute initialNameClass as nameClass
    While (nameClass has SuperClass)
    BEGIN
    Step 3: Submit SPARQL query with query user as object and find
    the SuperClass (subject) comparing similarity with predicates
    Step 4: Attach result
    END
Step 5: Submit SPARQL query with initialNameClass as subject and
find the geographical location (subject, predicate and object).
Step 6: Attach result
Step 7: Return result
End
```

We implemented this algorithm in a prototype using: Java, Eclipse Indigo (as IDE), Google Web Toolkit 2.5.1 to create a web client, Jena RDF framework to process (simplified) SPARQL queries and Virtuoso Server as triple store.

Figure 4 shows graphic interface to support user queries. We implemented a SPARQL Endpoint for INPA http://143.107.231.220:8890/sparql and implemented a set of queries described on experiments section.

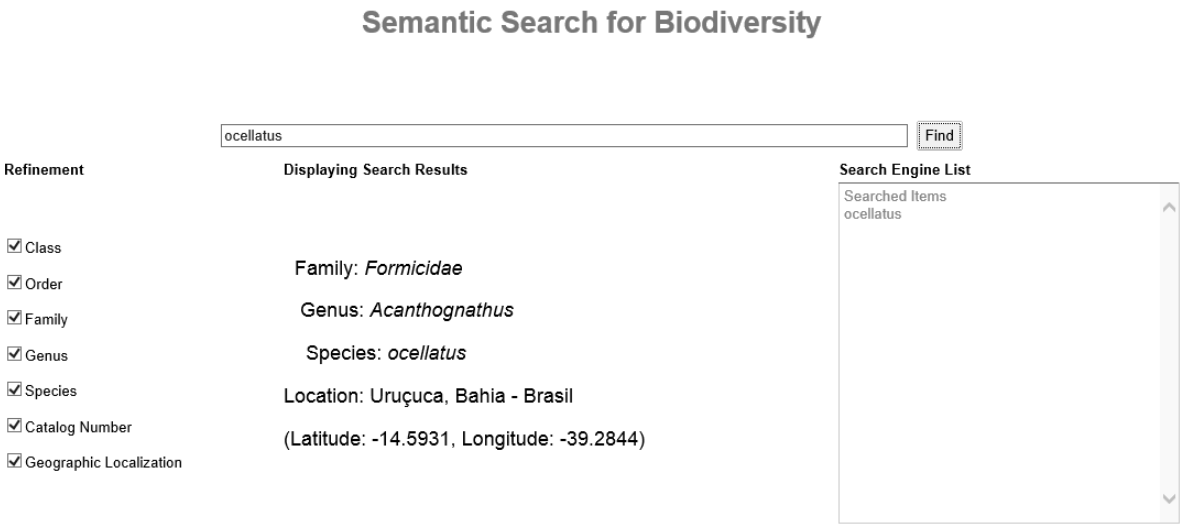

Figure 4. Web application for searching biodiversity information

5. Experiments

In order to validate our proposed architecture, researchers from our group and biodiversity scientists were interviewed to categorize important information from the INPA data.

We defined use cases (Table 1) with scenarios to identify the various user tasks and built SPARQL queries related with these use cases.

For each of the previous use cases, biodiversity experts identified the information set each user needed for each task and examples of queries that should have returned this information. After we tested each query, the same experts judged which results were relevant and non relevant (relevance non relevance judgment).

This process of information feedback is commonly referred to, in the literature, as relevance feedback [Salton 1971] when experts explicitly provide information on relevant documents to a query [Baeza-Yates and Ribeiro-Neto 1999]. In its original formulation, expert users inspect the query results and indicate those that are really relevant to the search. Table [tab:InfoNeeds] shows examples of users tasks and possible query strings to get the relevant biodiversity information.

Scientists can identify species using the taxonomic classification system no matter what their language. The taxonomic classification system is composed by a hierarchy (series of ranks) that shows the kinship of organisms and also, whenever possible, ancestor-descendant relationships.

Table 1. Biodiversity Use Cases

Use Cases	Goals	Queries
Use Case 01	Identification of a species.	Query1: fish ocellatus Query2: fish brasiliensis Query3: fish Corydoras splendens
Use Case 02	Determine information of a collect.	Query4: fish Hemigrammus gracilis Query5: Potamorrhaphis guianensis Query6: Hemigrammus guyanensis Query7: Iguanodectes spilurus
Use Case 03	Determine the best areas for aquaculture considering different types of species and geographical location of a collect.	Query8: Gnathocharax steindachneri

The basic ranks of the taxonomic classification system are kingdom, phylum, class, order, family, genus and species. The following SPARQL query (Listing 1) shows taxonomic system of classification for the *kingdom Animalia*.

Listing 1. SPARQL query returning the taxonomy of a specie

```
PREFIX oo: <http://www.owl-ontologies.com/
BiodiversityOntologyFull.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select ?phylum ?class ?order ?family ?genus ?species
where { oo:kingdom-Animalia oo:subkinofPhyKing ?phylum.
?phylum oo:subkinofClassPhy ?class .
?class oo:subkinofOrdClass ?order .
?order oo:subkinofFamOrd ?family .
?family oo:subkindOfGenFam ?genus .
?genus oo:subkindOfEspGen ?species .
}
```

The following SPARQL query (Listing 2) shows important information from a collect such as Collect, Research Institution, Method, Determinate Name.

Listing 2. SPARQL query returning information of a collect

```
PREFIX : <http://www.owl-ontologies.com/
BiodiversityOntologyFull.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select ?collect ?ResearchInstitution ?MethodCollect
?NameDeterminateCollect where {
?collect :mediationInstituicaoVinculo ?ResearchInstitution.
?collect :isClassifiedAsColetaTipoColeta ?MethodCollect .
?collect :mediationColetaRespColeta ?NameDeterminateCollect.}
```

The following SPARQL query (Listing 3) shows the geographical location of a specimen collect and other data, such as collect local, geographic space, latitude and longitude.

Listing 3. SPARQL query returning geographical location of a collect

```
PREFIX : <http://www.owl-ontologies.com/
BiodiversityOntologyFull.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select ?CollectLocal ?GeographicSpace ?latitude ?longitude
where {
?CollectLocal :localizationEspaGeoCoordGeo ?GeographicSpace.
?GeographicSpace :latitude ?latitude .
?GeographicSpace :longitude ?longitude .
}
```

To evaluate our semantic search architecture, we measured precision and recall to assess the performance of each approach dependent on input variable such as the user query. The recall value measures whether a tool retrieves all possible items related to the search terms contained in the data store, while precision measures to what extent only the relevant items were actually returned.

We compared the result in two search systems, our semantic search and keyword based search from SpeciesLink with data from INPA. We used a total of 16 queries (8 for each system).

To compare the results of only two systems, we will employ the Students T-tests, since they are designed for testing two data sets [B. Rasch and Naumann 2004]. When checking two data sets, each characterized by its average, standard deviation, and number of data points, it is possible to apply the T-test to identify, whether the means are in fact distinct or not. A probability value (p-values) below 0.05 indicates a statistically significant difference, whereas a p-value equal or exceeding 0.05 indicates no significant evidence, that there exists no significant difference between the performance values of two or more tools [Sachs 2003].

In our experiments, Semantic Search resulted in is significant difference in recall (p=0.0201 by t-test) and precision (p= 0.0006 by t-test) when compared to Keyword based search. One reason might be that keyword based search is not enough to capture the underlying semantics of user information needs, since it is content-oriented. This evaluation is shown in Table 2.

There is a significant difference in the mean of precision in Semantic Search minus the mean precision in Keyword Search equals 0.50416. The confidence interval of this difference from 0.302420283 to 0.705913042 is 95%. The mean of recall in Semantic Search minus the mean precision in Keyword Search is equals 0.20624745663. The confidence interval of this difference from 0.04330054489 to 0.36919436836 is 95%.

Table 2. Students T-tests

Group	Semantic Search (Recall)	Keyword Based Search (Recall)	Semantic Search (Precision)	Keyword Based Search (Precision)
Mean	0.587638057	0.3813906	0.975	0.470833338
Queries	8	8	8	8

6. Conclusions and Future Work

The architecture presented in this work provides a new document retrieval process by exploiting query terms to support scientists in the process of discovery and integrating biodiversity data and domain knowledge. This architecture can be classified, according to the categorization schema proposed by [Mangold 2007], as a Stand-Alone Search Engine. The search process uses resources labels from classes, properties, mappings and instances from domain ontologies represented in the OWL language.

We defined a mapping mechanism between relational database data and OntoBio ontology terms resulting in the generation of RDF triples (subject, object and predicate) saved in a triple store (Virtuoso). The triple stores make it much easier to add new predicates and write complicated queries or perform inferencing and rule processing.

A comparative analysis showed a significant increase in recall and precision in the semantic search. The possibility of creating queries that seek information based on relationships between data offers many alternatives to semantic search systems, since the results of these queries are not based only on specific information. Users can thus receive data that, in traditional systems, would not be considered by the query, but by analyzing their relations with other information, semantic search queries can consider them relevant.

As future work, we intend to extend our current implementation with more advanced structured searches in partnership with researches from INPA.

7. Acknowledgment

The authors would like to thank INPA for supporting this work. Thanks are also due to researchers of INPA's biological collections. This research was financed by the Brazilian funding agency CNPq.

References

Albuquerque, A. (2011). *Desenvolvimento de uma Ontologia de DomÃnio para Modelagem de Biodiversidade*. DissertaÃ§Ã£o de Mestrado. Universidade Federal do Amazonas.

B. Rasch, M. Friese, W. H. and Naumann, E. (2004). *Quantitative Methoden Band*. Springer, ISBN 978-3-540-33307-4.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Boley, H., Tabet, S., and Wagner, G. (2001). Design rationale of ruleml: A markup language for semantic web rules. pages 381–401.

GeoNames (2011). Geonames ontology. `http://www.geonames.org/ontology/documentation.html`. Accessed: 2013-07-30.

Kauppinen, T. and de Espindola, G. M. (2011). Linked Open Science—communicating, sharing and evaluating data, methods and results for executable papers. *Proceedings of the International Conference on Computational Science (ICCS 2011), Procedia Computer Science*, 4(0):726–731.

Latiri, C. C., Haddad, H., and Hamrouni, T. (2012). Towards an effective automatic query expansion process using an association rule mining approach. *J. Intell. Inf. Syst.*, 39(1):209–247.

Li, W. and Yang, C. (2008). A semantic search engine for spatial web portals. volume 2, pages II–1278 –II–1281.

Mangold, C. (2007). A survey and classification of semantic search approaches. *Int. J. Metadata Semant. Ontologies*, 2(1):23–34.

Mariano R, M. and Calvanese, D. (2012). *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.

Mittal, N., Nayak, R., Govil, M. C., and Jain, K. (2010). A hybrid approach of personalized web information retrieval. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 308–313.

PATO (2010). The phenotypic quality ontology. `http://bioportal.bioontology.org/ontologies/1069`. Accessed: 2013-07-30.

Sachs, L. (2003). *Angewandte Statistik: Anwendung statistischer Methoden*. Springer, November. ISBN 3540405550.

Salton, G., editor (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey.

Santos, V. D., Baiao, F. A., and Tanaka, A. (2011). An architecture to support information sources discovery through semantic search. In *Information Reuse and Integration*.

WGS84 (2003). W3C Semantic Web Interest Group: Basic Geo (WGS84 lat/long) Vocabulary.

Xiong, J., Huang, W., and Jin, C. (2009). An ontology-based semantic search approach for geosciences. volume 3, pages 87 –90.