Improving biodiversity data retrieval through semantic search and ontologies

IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2014, Warsaw.
http://www.producao.usp.br/handle/BDPI/48658

Improving Biodiversity Data Retrieval through Semantic Search and Ontologies

Flor K. Amanqui[1], Kleberson. J. Serique[1], Silvio. D. Cardoso[1]
José. L. Dos Santos[2], Andrea. Albuquerque[2], Dilvan A. Moreira[1]
[1]University of São Paulo, ICMC, São Carlos, Brazil
[2]National Institute for Amazonian Research, Manaus, Brazil
Email: {flork, serique, dilvan}@icmc.usp.br, silvio.domingos.cardoso@usp.br
{lcampos, andrea}@inpa.gov.br

***Abstract*—Due to the increased amount of available biodiversity data, many biodiversity research institutions are now making their databases openly available on the web. Researchers in the field use this databases to extract new knowledge and also share their own discoveries. However, when these researchers need to find relevant information in the data, they still rely on the traditional search approach, based on text matching, that is not appropriate to be used in these large amounts of heterogeneous biodiversity's data, leading to search results with low precision and recall.**

We present a new architecture that tackle this problem using a semantic search system for biodiversity data. Semantic search aims to improve search accuracy by using ontologies to understand user objectives and the contextual meaning of terms used in the search to generate more relevant results. Biodiversity data is mapped to terms from relevant ontologies, such as Darwin Core, DBpedia, Ontobio and Catalogue of Life, stored using semantic web formats and queried using semantic web tools (such as triple stores). A prototype semantic search tool was successfully implemented and evaluated by users from the National Research Institute for the Amazon (INPA). Our results show that the semantic search approach has a better precision (28% improvement) and recall (25% improvement) when compared to keyword based search, when used in a big set of representative biodiversity data (206,000 records) from INPA and the Emilio Gueldi Museum in Pará (MPEG). We also show that, because the biodiversity data is now in semantic web format and mapped to ontology terms, it is easy to enhance it with information from other sources, an example using deforestation data (from the National Institute of Space Research - INPE) to enrich collection data is shown.

***Index Terms*—Semantic Web; Semantic Search; Ontology; Data Integration; Biodiversity.**

I. Introduction

Nowadays, the Web has become one of the main sources of biodiversity information. Biodiversity research institutions continually add new specimens and their related information to their biological collections and make this information available on the Web. These collections provide, among other things, detailed information about specimens distribution in space and time. Most specimen information indicates where the item was located, when it was collected and by whom [1]. This information about location of specimen occurrences can be combined with other geo-referenced data for predictive distribution maps.

Even though the potential impact of using collection data with other geo-referenced data is enormous, the huge data volume of these collections, which continues to grow, is a difficult obstacle. Responding to the global interest in biodiversity conservation and sustainable development, several projects are under way to digitize important worldwide biodiversity collections. Some of these projects are: the Global Biodiversity Information Facility[1] (GBIF), the Biodiversity Database Collection of the National Research Institute for the Amazon (INPA)[2], Large-Scale Biosphere-Atmosphere Experiment in Amazonia[3] (LBA), Reference Center on Environmental Information[4] (CRIA), and The New York Botanical Garden[5] (NYGB). Many other projects exist with a mix of regional and/or specific aims. However, these projects do not have a standard or automatic way to represent their data and do not interoperate.

To find relevant data, from the huge amount of biodiversity information present on the Web, an efficient searching architecture is required. Search engines could be a solution to this problem of finding relevant biodiversity information from different sources. A search engine algorithm is based on trying to match keywords, from a user's keyword list, to strings from indexed records (e.g. from web pages) to generate a ranked list of search results.

Although search engines are very helpful in finding information on the Web (and get smarter all the time), they suffer from the fact that they do not know the meaning of the terms and expressions used in Web pages (or other kinds of records) and the relationships between them. In the biodiversity field, it is not different. The large quantity of data generated by research institutions is difficult to search. To overcome the

[1]http://www.gbif.org
[2]http://colecoes.inpa.gov.br
[3]http://lba.inpa.gov.br
[4]http://www.cria.org.br
[5]http://www.nybg.org

978-1-4799-4143-8/14 $31.00 © 2014 IEEE
DOI 10.1109/WI-IAT.2014.44

search engine problems and to be able to retrieve relevant and meaningful information intelligently, the Semantic Web was proposed by [2].

The Semantic Web is considered the new-generation of the Web that tries to represent information in such a way that it can be used by machines, not just for display purposes, but for automation, integration and reuse across applications [3]. The key idea is to add semantics into the Web content in order to make it easier to find and use for both humans and machines.

The next generation of the Semantic Web promises to increase the performances and the relevance of search engines, by first attaching formal semantics to resources, and then exploiting this semantics during the search process [4]. According to [5], the semantic search approach tries to augment and improve searches on a set of resources that are initially unknown to the user, by using ontologies and semantic annotations of these resources. Also, semantic search aims to improve search accuracy by understanding user objectives and the contextual meaning of the terms used in the search, as they appear in the searchable data, either on the Web or within a closed system, to generate more relevant results.

We present a new architecture that uses a semantic search system for biodiversity data and semantic web formats and tools to represent this data. It supports mapping between biodiversity data and the ontologies describing it. A prototype based on this architecture was implemented.

This prototype was tested using a set of representative data about biodiversity (206,000 records) from the National Institute for Amazonian Research (INPA) and Emilio Gueldi Museum in Pará (MPEG), two of the most important institutions doing research in biodiversity in the Amazon Forest. This data was downloaded from the SpeciesLink web site[6]. SpeciesLink is a distributed information system that integrates primary data from biological collections from many research institutions from Brazil and abroad. It is also a popular online tool to search for biodiversity data. The test results showed a 28% improvement in precision and 25% in recall, when comparing our semantic search approach to keyword based search using the SpeciesLink search tool.

We also show easy data interoperability with other open data sources, which also use semantic web formats and ontology terms, through an example using deforestation data (from the National Institute of Space Research - INPE) to enrich collection data.

The remainder of this article proceeds as follows: Section II discusses related work. Section III shows the architecture for semantic search. Section IV presents a synopsis of our experiments results and Section V concludes by summarizing our results and describing future works.

II. Related Work

We studied a number of techniques for biodiversity information retrieval based in keyword based search and semantic search. The techniques for keyword based search basically determine which collection records contain the keywords in the user query [6]. A survey of the available literature indicates limitations in keyword based search techniques:

- According to [7], the search concentrates on the keyword matching of user query with indexed documents, while ignoring the semantic of the query. A term may have several synonyms that are not considered while returning the search results to the user due to their unavailability.
- Words used by users can have problems, such as synonym or words with many meanings, that are very difficult to solve. People often choose keywords subjectively, arbitrarily and lacking standardization. Information retrieval based on keywords (at the syntax level) focus on simply matching keywords, without the ability of knowledge representation, processing and understanding [8].
- According to [9], keyword-based search is not sufficient to capture the underlying semantics of user information needs, since it is content-oriented.

Even though keyword-based search have all this limitations, it is still the main and, in most cases, the only search tool available in major biodiversity repositories, such as:

- GBIF Data Portal[7] is a service that provides access to millions of scientific data records about biodiversity that are being shared via the Global Biodiversity Information Facility (GBIF) network. In March 2014 there were 405,720,566 data records (352,593,699 georeferenced) accessible from this portal.
- SpeciesLink[8] is a distributed information system that integrates primary data from biological collections from diverse institutions, such as museums, herbaria and microbiological collections, from Brazil and abroad. It had, in March 2014, 326 collections and sub-collections 6,425,366 on-line records (2,719,146 georeferenced).

New search approaches have been proposed to overcome the terminology and meaning mismatch limitations in keyword based search. A number of techniques have been developed for using ontologies to retrieve relevant documents in response to a query. We list the ones we considered most related to the biodiversity field:

- In [10], a semantic search approach for geosciences is proposed in which a query agent is linguistically mapping lexicon vocabularies to concepts and relationships from geological ontologies.
- In [11], a semantic search system for ecology data is presented. It allows structured searches over user annotations using ontology terms. Authors have used the Extensible Observation Ontology (OBOE) for query expansion.
- Fedel et.al. [12] present the specification and implementation of a framework to process multimodal queries that support both text and images as search parameters for biodiversity studies. This framework extends queries on observation data by combining standard text-based queries with ontology manipulation and query by image content.

These systems use relational databases to store the biodiversity data and ontologies. The data being used, in each one, has to

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

[7]http://www.gbif.org/

[8]http://splink.cria.org. br

use the system's database schema forming a closed system. Most of the search techniques used require complex analysis, involving natural language processing, to discover the implicit context and semantics of query terms in relational databases (what is a limiting factor). Data stored in one system cannot be queried from another. Third part applications cannot easily query or share the data using, for instance, Linked Open Data (LOD) technologies.

III. SEMANTIC SEARCH ARCHITECTURE

This section presents our semantic search architecture for biodiversity data. The development of this architecture was divided in three parts:

1) The Biodiversity Ontology, which play a central role in our semantic search architecture by providing a shared knowledge,
2) The Mapping Component, which maps the collection records to ontology entities,
3) The Web Interface, which process queries from either users, using a web interface, or machines, using a SPARQL Endpoint.

Figure 1 presents the architecture's overall schema.

A. The Biodiversity Ontology

Among Semantic Web technologies, ontologies play a central role by providing a shared knowledge about the objects in the real world. They promote reusability and interoperability among different sources [8]. To deal with biodiversity data, we modified a biodiversity ontology (OntoBio) that is utilized to associate semantic meaning (terms) to data.

OntoBio was designed to conceptualize knowledge about biodiversity collection data. Originally created by INPA [13], it is being jointly developed by it and us (at ICMC - University of São Paulo). Its main objective is to provide a clear and precise conceptualization of the information describing specimen's collections. Ontobio is divided into five sub-ontologies (Collection, Material Entity, Spatial Location, Ecosystem, and Environment), integrated by relationships between their concepts and axioms (Figure 2). OntoBio was modeled using the OntoUML language, as its formal language for conceptual modeling, allowing it to capture complex aspects of the biodiversity domain. The development of OntoBio is only being possible due to the help of the highly capable experts, from INPA, willing to contribute to the project. The complete ontology is presented in details in [13].

One of the advantages of having data annotated using OntoBio concepts (for that matter, using any open ontology) is that it can be reused as Linked Open Data (LOD). LOD describes a method of publishing structured data so that it can be interlinked and become more useful [14]. To better archive that, data annotated using OntoBio has to be easily interlinked with other data already available on the web (as part of the wider LOD community) through the use of as many shared concepts as possible. With that in mind, we rewrote the original version of OntoBio to reuse, whenever possible, terms from other public available ontologies to allow better "linkability" with data already annotated using them.

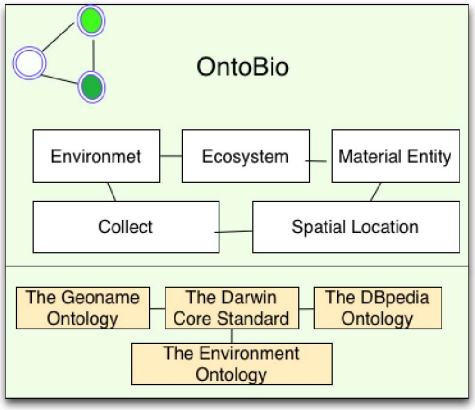

Figure 2: Overview of the biodiversity ontology (OntoBio).

When reusing an element from another ontology, we copied its URI and any axioms related to it that we needed. Then, if necessary, we added new axioms to it. We added terms from the following public ontologies or controlled vocabularies (all available in the OWL or RDF languages):

- The Environment Ontology[9] (EnvO), which provides a controlled, structured vocabulary that is designed to support the annotation of any organism or biological sample with environment descriptors. EnvO contains terms for biomes, environmental features, and environmental material. In OntoBio, we use it to describe biomes (`ENVO:00000428`) and other environmental features. Examples of biome terms are: boreal moist forest biome, tropical rainforest biome, and oceanic pelagic zone biome. EnvO is available to view or download in the Bioportal public Web site. The BioPortal is a Web portal that provides access to a library of biomedical ontologies and terminologies via the National Center for Biomedical Ontology (NCBO) Web services.
- The Darwin Core Standard[10] includes a glossary of terms intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. In OntoBio, we use it to describe properties, elements, fields, columns, attributes and concepts.
- The Geonames Ontology[11], which makes it possible to add geospatial semantic information to the Word Wide Web. Over 8.3 million geonames toponyms now have a unique URL with a corresponding RDF web service. Other services describe the relation between toponyms. The GeoNames Ontology is available in OWL.
- The DBpedia Ontology, which is a community-curated ontology consisting of 320 classes which form a subsumption hierarchy and are described by 1,650 different properties. The ontology is maintained and extended by the community in the DBpedia Mappings Wiki. This community also creates mappings from Wikipedia

[9]http://bioportal.bioontology.org/
[10]http://www.tdwg.org/standards/
[11]http://www.geonames.org/ontology/

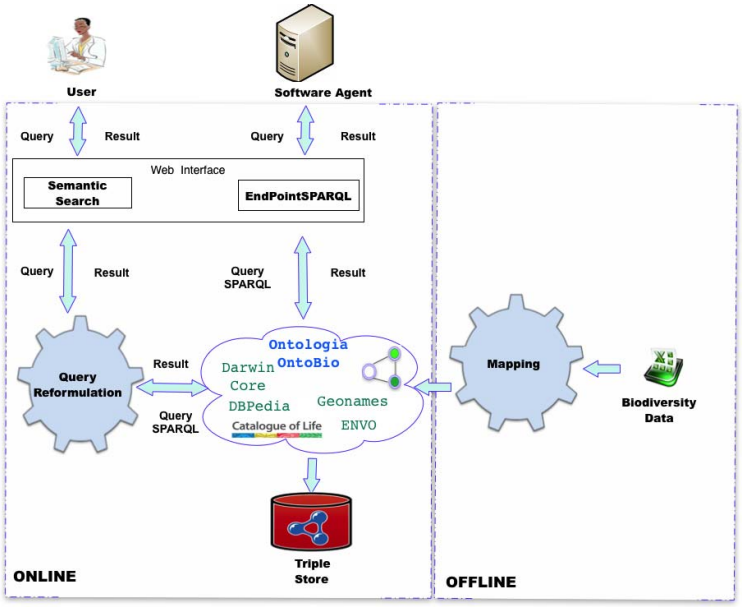

Figure 1: The Architecture for a Semantic Search enabled biodiversity data repository

information representation structures (info-boxes created by Wikipedia editors) to the DBpedia ontology. These mappings are used to automatically create instances of the ontology from Wikipedia information (in Wikipedia's many languages), which ensures a huge coverage of topics. OntoBio uses instances from the English (dbpedia.org) and Portuguese (pt.dbpedia.org) mappings of DBpedia. The Portuguese mapping is mainly used used to describe Brazilian cities, such as `http://pt.dbpedia.org/resource/Urucurituba`.

The Protégé 4 ontology editor was used to write the new OntoBio ontology version in OWL 2 DL. This new version has a dereferenceable URI[12], meaning that the URI can be used by tools, e.g. Protégé 4, to get the ontology automatically from the web (as an OWL file).

1) Species Taxonomy: In addition to OntoBio, we need a biological taxonomy to classify species. A taxonomy is just an ontology where each class can have just one parent. Unfortunately, there is no standard taxonomy used by all biologists. Biodiversity experts from INPA recommended the use of the taxonomy created by the Catalogue of Life[13], an online database of the world's known species of animals, plants, fungi and micro-organisms (with 1.35 million species). The Catalogue of Life taxonomy is not available for download as a separate file in RDFS or OWL. So we had to write a program to use the site web services[14] to navigate through its taxonomic tree and write it as an OWL file.

B. The Mapping Component

The Mapping Component loads the domain ontologies, taxonomic information and the biodiversity data collection to transform them in a set of RDF triples. We used a small Domain Specific Language (DSL) to represent the mapping between rows of data tables into OntoBio classes and properties, to create RDF triples.

Data from INPA and MPEG, and from dozens of other biodiversity research institutions, is available in the SpeciesLink web site. SpeciesLink offers this data in csv text files using a format based on Darwin Core. We used the mapping component to convert all INPA's and MPEG 's records for the Brazilian state of Amazonas from the SpeciesLink web site to RDF triples. This mapping is done offline and generates the triples that will be stored in the triple store (in our case, Virtuoso) and queried during user searches.

Virtuoso is an open source triple store with very good performance. It is used in sites like the DBpedia SPARQL Endpoint. The Virtuoso also works with multiple RDF graphs (knowledge trees) at the same time and supports the SPARQL 1.1 query language. It also provides a faceted browser user interface for querying the RDF data store.

This mapping is illustrated in the Algorithm 1.

This algorithm is capable of:

(i) Create an ontology individual (entity) representing each specimen in the collection.

(ii) Automatically link specimen name to taxon data using the Catalogue of Life[15] (CoL) webservices. Each collection record receives a URI connecting it to a taxon id in the CoL website;

(iii) Automatically link geographic information to the DBpedia, the Wikipedia Linked Data version. For instance, the DBpedia URI for each city is added to the record of each specimen collected;

(iv) Convert strings representing dates in various formats to

[12]http://purl.org/biodiv/ontobio

[13]http://www.catalogueoflife.org

[14]description at http://webservice.catalogueoflife.org

[15]http://www.catalogueoflife.org/

Algorithm 1 Algorithm for mapping biodiversity data onto ontology terms (OWL individuals).

Input: Collection data in csv format
Output: Owl file with ontology and triples
for all $rows$ in $Collection$ **do**
  Get species name of specimen
  Use CoL webservices to link name to taxonomy id
  if $species\ name\ not\ found$ **then**
    Try next taxonomic rank until CoL can find an id
  end if
  Create an OWL individual to represent specimen
  Assert that it belongs to taxon returned by CoL
  Add label with taxon name
  Get municipality of colection
  Use DBpedia webservices to find municipality and state URIs
  Add municipality and state URIs
  if $latitude\ and\ longitude\ available$ **then**
    Convert them to proper format
    Add them to OWL individual
  end if
  Add OWLLiterals for colection and classification dates
  Add other information, such as gender, institution, etc
  Check for semantic errors in individual
end for
Dump OWL file to triple store
return

the proper equivalent XSD date type (used by RDF) and check for semantic errors, such as an animal species being declared as belonging to the plant kingdom.

After mapping the data from the repository into RDF triples, they are loaded into a triple store (in this case, Virtuoso) and used by the Web Interface to do its SPARQL queries.

C. Web Interface

The Web Interface is responsible for the interaction between users and our semantic search engine. The search process begins with an initial keyword list entered by a user (biodiversity specialist) that represents his/her search intentions. When typing these keywords, a widget from BioPortal (called *term-selection field*) suggests new terms based on semantic expansion and dictionary similarity with terms from BioPortal hosted ontologies (OntoBio included). Users may or may not accept the suggested terms. The search results display consists of a vertical list of document titles and several lines from the records that fulfill the search criteria (family, genus, specie and other information). The user interface is shown in Figure 3:

The queries entered in the Web interface are processed by the query reformulation component.

D. The Query Reformulation Component

The Query Reformulation is done online and receives the search terms from the user, it finds all classes of ontologies to which the term belongs and expands the keyword list using an

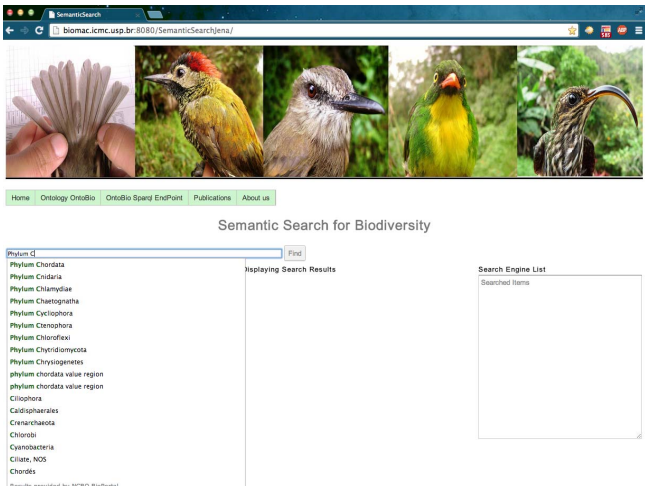

Figure 3: Web application for searching biodiversity information.

Algorithm 2 Semantic Search algorithm developed to compare ontology terms, derived from user input keywords, to the OntoBio graph in a triple store.

Input: $userQuery$ a set of words typed by user
Output: $results$ a information list from the triples
Connect to Triple Store and OntoBio ontology
for all $word$ in $userQuery$ **do**
  submit a SPARQL query with $word$ as subject
  add return to $results$ {$word$?predicate ?object}
  submit a SPARQL query with $word$ as predicate
  add return to $results$ {?subject $word$?object}
  submit a SPARQL query with $word$ as object
  add return to $results$ {?subject ?predicate $word$}
end for
for all unique $objects$ and $subjects$ in $results$ **do**
  submit a SPARQL query to find its location
  add return to $results$ {:longitude and :latitude}
end for
return $results$

Algorithm 2. The basic idea of this algorithm is to compare ontology terms, derived from user input keywords, to the OntoBio graph in the Virtuoso triple store.

When a user submits a query, each word in the query is compared with the labels from the triples in the Virtuoso triple store (using SPARQL queries). This includes OntoBio terms and collection data. All relevant triples are found.

If a triple also has latitude and longitude data attached to it, these are also collected. The algorithm results, obtained using multiple SPARQL queries, are presented to the user in the Web Interface.

This algorithm was implemented in a prototype, a Rich Internet Application (RIA), using the Java technology. The server side used the Jena RDF framework, to reason about SPARQL queries, in addition to a Virtuoso triple store. The

Web Interface was implemented using Google Web Toolkit (GWT 2.6) on the client side.

IV. Experiments

In order to validate our architecture and guide the prototype tests, INPA's biodiversity experts were interviewed to categorize important information from INPA's and MPEG's data (e.g. genus, family, species, description of location, etc.).

These interviews helped us to understand more about their work and to form a common ground for discussions. Because this technology is so new, it is difficult for us and our partners at INPA to foresee all its possible uses. To help us, we defined use cases with features and scenarios to identify the various user tasks. One such use case is presented below:

USE CASE 01: Classification of Ecologically Degraded Areas

USER: Christine Smith, 32 years-old, biologist, NGO employee.

GOAL: To determine if areas in the state of Pará, Brazil, are ecologically degraded based on the size of their deforested areas and species collected there.

MOTIVATION: The presence or not of some species of plants and animals can serve as biological markers (bio indicators) that indicate the degree of conservation or degradation in a habitat.

TASKS

(i) Find deforestation information: The "Linked Brazilian Amazon Rain Forest Data" SPARQL EndPoint divides the Brazilian forest in 25 km² squares with deforestation information.
(ii) Link the geographic information of collected specimens to their deforestation level.
(iii) Use the information to plot maps using tools such as the R language (software environment for statistical computing and graphics).

NECESSARY TOOL FEATURES

(i) Specification of bio-marker species using the species name or any higher taxonomic level, like phylum, genus or family.
(ii) EndPoint SPARQL to answer queries from computer programs, such as the a R script, to allow data integration with other repositories.
(iii) Flexible way to limit habitats of interest (one species can be a bio-marker in one habitat but not in another). For example, Christine can specify an aquatic habitat for an insect without having to specify if the this habitat is a river, stream or lake.

After studying our use cases, we divide the experimental evaluation of our prototype in three parts: creation of triples using INPA's and MPEG's insect and fish collections, semantic search testing and linking with deforestation data.

A. Creation of Triples from insect and fish collections

The first experiment demonstrated the mapping mechanism, using the OntoBio ontology and INPA's and MPEG's collections data, to obtain a set of triples (subject, object and predicate).

We used the Algorithm 1 to convert all INPA's and MPEG's records for the Brazilian state of Amazonas from the SpeciesLink web site (206,000 records) to RDF triples. This RDF data was stored in our Virtuoso Triple Store and can be explored using SPARQL queries. The mapping program was able to treat defective input records, such records lacking fields.

The experiment demonstrated that the mapping method works even when the data being used has defective records. To integrate the biodiversity data in RDF to the wider LOD data community on the web, we setup a SPARQL EndPoint[16]. Our EndPoint allows that third part programs query our knowledge base, via the SPARQL language, and reuse it in their applications.

B. Semantic Search Testing

For the previous use case, biodiversity experts from INPA identified the information set (from the data converted to RDF triples) that each user needed for each task. They also created queries examples 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 information feedback process is commonly referred to, in the literature, as relevance feedback [15]. In its original formulation, expert users inspect the query results and indicate those that are really relevant to the search. Table I shows examples of user tasks and possible query strings to get the relevant biodiversity information.

Table I: User tasks and query examples.

User Tasks	Queries
Differentiate scientific name. The scientific name is what scientists sometimes call something to differentiate it from other things in the same order/family/genus. For example, there are insects that have the same scientific name as fish.	fish ocellatus, fish brasiliensis, Corydoras guianensis, steindachneri Crenicichla, Geophagus
Determine if group members breed with organisms outside their group. A difficult task for biologists is to determine if a group of organisms is a separate species.	Hypsiboas, Anolis, hoplias, camponotus jupiaba
Determine the geographic distribution of specimens. A common task for biologists is the monitoring of wildlife population levels.	nannostomus bryconops, Serrasalmus acestrorhynchus

Researchers from INPA and system analysts from USP participated in the experiments. A total of 28 distinct queries were done. It is important to remember that it takes a lot of effort, from the domain experts, to figure out which data should be returned in each query, given the size of the data set, and the relevance of the returned results. We tested and compared the result of two search systems: our semantic search engine and the keyword based search system used by the SpeciesLink site (this site uses a standard search system for biodiversity data), both using data from INPA and MPEG.

The evaluation of the two search approaches, using the average precision, recall and F1 is shown in Figure 4. The F1 score can be interpreted as a weighted average of the precision and recall scores. A F1 score reaches its best value at 1 and worst at 0.

[16]http://biobak.icmc.usp.br:8890/sparql

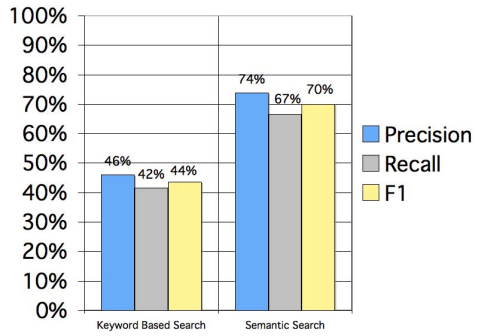

Figure 4: Precision, Recall and F1 Experimental Results.

Our semantic search architecture reached a highest precision around 28% (on average) better than using the keyword search, in the scenarios described by the biodiversity experts.

The semantic search recall was 25% (on average) better than the keyword based search (from SpeciesLink). This results demonstrate that the values for precision and recall, for the semantic search, were significantly better.

In all use case scenarios, the semantic search obtained better F1 measure results (26% better on average than the keyword search), reflecting a good balance between the increased precision and recall obtained.

Precision, recall, and the F1 are set-based measures. We extended these measures to evaluate the ranked retrieval results in the Semantic Search and Keyword Based Search. The traditional way of doing this is the 11-point interpolated average precision. For each such set, precision and recall values can be plotted to give a precision-recall curve considering the top k =10 retrieved documents, such as the one shown in Figure 5.

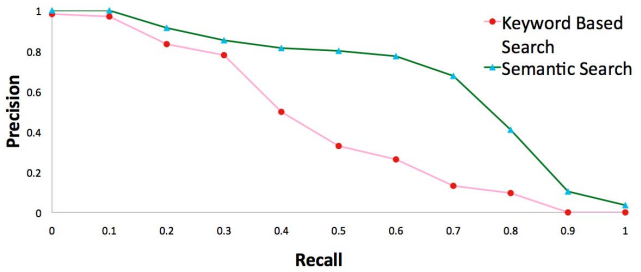

Figure 5: Averaged 11-point precision/recall graph across 28 queries for a representative semantic search and keyword based search system.

It is very important to point out that the data, shown in Figure 5, only takes into account cases where the keyword based search actually did return something. If the keywords are not present in the records, methods based only on keyword search fail. As an example, if users try to find plants records of the phylum "tracheophyta" in the INPA's dataset in the SpeciesLink site (keyword search), they will find none. The problem is that the phylum field in this dataset is empty. Using the semantic search, as shown in the next section, to search for "tracheophytas" would result in 61244 records. The difference is that species, in the semantic search, are elements of a taxonomy. In this way, independently of a species having all its taxonomic levels stated in the INPA's records, the search engine can find specimens belonging to it based in any taxonomic level. More over, the INPA's records only have seven taxonomic levels (kingdom, phylum, class, order, family, genus and species), if users use other levels, for instance super family, keyword search fails while semantic search does not.

C. Linking to Amazon Deforestation Data

Semantic or keyword based search can be useful to peruse data (to find out if it is useful for some application) or to find a small number of specific records. But when it comes to data analysis, computer programs are necessary. There are two ways to write these computer programs:

- Applications developed by programmers for use by biologists.
- Scripts written by biologists in some DSL.

Lets focus on the second case, it is simpler and does not need any programmers to work.

For this test, we downloaded all INPA's records (August 2013) for the state of Amazonas - Brazil (108,220 in total) and, using a script and a mapping (written in a small DSL), read all plant records (53,120 in total) and transformed them into RDF triples using the OntoBio ontology. These triples were loaded and made available in our Virtuoso triple store. Finally we chose the "Linked Brazilian Amazon Rainforest" dataset [14] to extract information to connect to our dataset. This dataset consists of 8250 cells—each of size of 25 km by 25 km—capturing the observations of deforestation in the Brazilian Amazon Rainforest and a number of related and relevant variables. It is available in a SPARQL EndPoint [17] and via dereferenceable URIs.

With the help of R, a DSL language popular among biodiversity researches, and its packages *sparql* and *so* we created a script to:

(i) read, using a SPARQL query, all the cells belonging to the Amazonas state from the "Linked Brazilian Amazon Rainforest" dataset with their positions (a polygon) and deforestation percentage (2008 values).
(ii) read, using a SPARQL query, all occurrences containing plant samples belonging to the phylum Tracheophyta that had latitude and longitude information (19,887) from our INPA's dataset.
(iii) for each cell from step 1, which had a deforestation ratio smaller than 10%, count the number of collections made inside it.
(iv) plot the cells with colors representing the number of collections in each.

Figure 6 shows the generated map. The brown color shows areas with deforestation bigger than 10% (no specimens counted). The map shows a fairly amount of specimens coming from a small area. The black square has the biggest number of collections, 835, and is far greater than the next one, with

[17] at http://spatial.linkedscience.org/sparql

380. We separated it to get a better color distribution in the map.

The map shows that we can merge our data with other Linked Data resource to enhance its informative value. Using the results, users can know which collections come from areas there are still preserved and see their distribution in a map. One of the big advantages of having INPA's data in RDF is to be able to connect it to other sources.

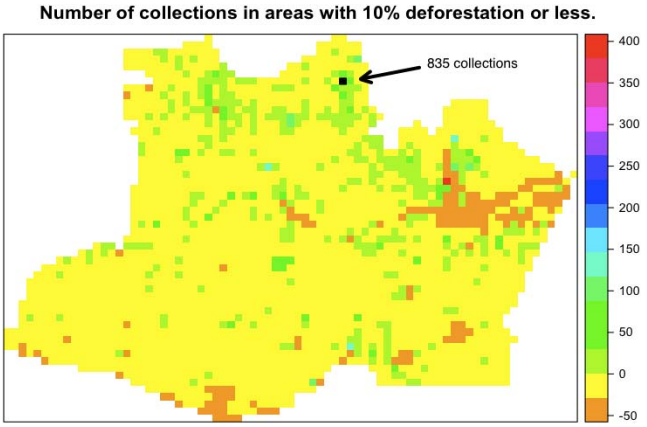

Figure 6: Areas where plant samples of the phylum Tracheophyta where collected in the state of Amazonas (Brazil). The color brown indicates areas with more than 10% deforestation (2008 data), samples from them were not counted. The distribution is quite concentrated in a few areas, the black spot represents the area with more collections (830).

V. Conclusions and Future Work

We presented a new architecture that uses a semantic search system to query biodiversity data and semantic web formats and tools to represent it. We implemented a prototype of it and, with it, showed that the architecture produces results with better precision, recall and F1-measure than a standard biodiversity search tool (SpeciesLink) using real biodiversity data from INPA and MPEG collections.

We demonstrated that, once collection data is in RDF format, it is easy to integrate it with data from different and independent data sources (if they share common ontology terms). We also showed that it is possible, for a biodiversity expert, to write a DSL script himself to do this data integration and answer scientific questions (using R, a popular scripting language among biologists). Data integration and scripting are very important because, in many cases, after finding relevant data, it is almost impossible to manipulate it by hand given its size. So, experts have to write scripts to integrate and analyze data to get the answers to their scientific questions.

As future work, we plan to refine our use cases. We intend to reuse geoSPARQL[18] ontology terms to describe georeferenced data, such as shapes of municipalities, national parks and farms, to include geographical semantic information in queries. We also intend to extend our current implementation with more advanced structured search types, in partnership with INPA's researchers.

[18]http://www.opengeospatial.org/standards/geosparql

Acknowledgments

The authors would like to thank INPA for supporting this work and, in special, their biological collections experts and professor . This research was financed by the Brazilian funding agency CNPq.

References

[1] J. L. C. dos Santos, "A biodiversity information system in an open data/metadatabase architecture," Ph.D. dissertation, Enschede.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, May 2001. [Online]. Available: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[3] H. Boley, S. Tabet, and G. Wagner, "Design rationale of ruleml: A markup language for semantic web rules," 2001, pp. 381–401.

[4] A. D. Miron, J. Gensel, M. Villanova-Oliver, and H. Martin, "Towards the geo-spatial querying of the semantic web with ontoast," in *Web and Wireless Geographical Information Systems, 7th International Symposium, W2GIS 2007, Cardiff, UK, November 28-29, 2007. Proceedings*, ser. Lecture Notes in Computer Science, J. M. Ware and G. E. Taylor, Eds., vol. 4857. Springer, 2007, pp. 121–136.

[5] C. Mangold, "A survey and classification of semantic search approaches," *Int. J. Metadata Semant. Ontologies*, vol. 2, no. 1, pp. 23–34, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1504/IJMSO.2007.015073

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

[7] A. K. Sharma, N. Duhan, and B. Sharma, "A semantic search system using query definitions," in *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia*, ser. IITM '10. New York, NY, USA: ACM, 2010, pp. 279–283. [Online]. Available: http://doi.acm.org/10.1145/1963564.1963613

[8] C. Zhao, J. Wang, W. Hu, X. Yu, and X. Wang, "An ontology-based semantic search model study," Oct. 2010, pp. 182 –185.

[9] V. Dos Santos, F. Baiao, and A. Tanaka, "An architecture to support information sources discovery through semantic search," Aug. 2011, pp. 276 –282.

[10] J. Xiong, W. Huang, and C. Jin, "An ontology-based semantic search approach for geosciences," vol. 3, Dec. 2009, pp. 87 –90.

[11] C. Berkley, S. Bowers, M. Jones, J. Madin, and M. Schildhauer, "Improving data discovery for metadata repositories through semantic search," Mar. 2009, pp. 1152 –1159.

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

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

[14] T. Kauppinen and G. M. de Espindola, "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*, vol. 4, no. 0, pp. 726–731, 2011.

[15] G. Salton, Ed., *The SMART Retrieval System - Experiments in Automatic Document Processing*. Englewood, Cliffs, New Jersey: Prentice Hall, 1971.