Semantics for Biodiversity (S4BioDiv 2013)

Proceedings of the First International Workshop on Semantics for Biodiversity

May 27th 2013
Montpellier, France

Pierre Larmande,
Elizabeth Arnaud,
Isabelle Mougenot,
Clement Jonquet,
Thérèse Libourel,
Manuel Ruiz
(editors)

May 2013
Photo credit @ I. Mougenot
Logo credit @ P. Larmande

Semantics for Biodiversity (S4BioDiv 2013)

Proceedings of the First International Workshop on Semantics for Biodiversity

Montpellier, France, May 27, 2013

In conjunction with ESWC 2013 (http://2013.eswc-conferences.org)

Edited by

Pierre Larmande *
Elizabeth Arnaud **
Isabelle Mougenot ***
Clement Jonquet ****
Therese Libourel ***
Manuel Ruiz *****

* IRD - UMR DIADE, France
** Bioversity International, Montpellier, France
*** Université Montpellier 2 - UMR Espace-Dev, France
**** Université Montpellier 2 - LIRMM, France
***** CIRAD - UMR AGAP, France

Web site

http://semantic-biodiversity.mpl.ird.fr

On line proceedings

CEUR Workshop Proceedings, volume 979

http://ceur-ws.org/Vol-979

Organization Committee

Pierre Larmande (IRD - UMR DIADE, France), **Organization Chair**

Isabelle Mougenot (UMII - UMR Espace DEV, France) co-Chair

Thérèse Libourel (UMII - UMR Espace DEV, France)

Clément Jonquet (UMII - LIRMM, France)

Scientific Program Committee

Elizabeth Arnaud (Biodiversity International), **PC Chair**

Manuel Ruiz (CIRAD - UMR AGAP, France)

Pierre Bonnet (CIRAD - UMR AMAP, France)

Richard Bruskiewich (Bioversity International)

Pascal Neveu (INRA - UMR MISTEA, France)

Joel Sachs (University of Maryland, Baltimore County,USA)

Julie Chabalier (Natural Solutions, France)

Cyril Pommier (INRA - URGI - France)

Mark Wilkinson (Polytechnic University of Madrid, Spain and University of British Columbia)

Konstantin Todorov (UMII - UMR LIRMM, France)

Pankaj Jaiswal (Oregon State University,USA)

Xavier Sirault (CSIRO, Autralia)

Caterina Caracciolo (FAO , Italy)

Lieke Verhelst (Wageningen University, The Netherlands)

Damian Gessler (iPlant Collaborative, USA)

Nikos Manouselis (Agro-Know, University of Alcala, Spain)

Eamonn O Tuama (GBIF, University of Copenhagen , Denmark)

Norman Morrison (NEBC, The University of Manchester, UK)

Foreword

Semantic web standards, tools, ontologies and related technologies have considerably matured in the recent years. Nowadays, accessing a wide catalogue of biological, social, environmental, and ecological data sources helps stakeholders working on biodiversity to answer their complex questions. Will the real time access to web resources effectively support the definition of strategies to conserve and manage biodiversity? How might semantic web technologies help us to handle the complex and heterogeneous big data related to biodiversity?

The workshop aims to identify the key challenges faced by the bioinformatics community, discuss potential solutions and identify the opportunities emerging from the trans-disciplinary interactions between Plant Science and Informatics experts. Therefore, we expect the bioinformatics experts to explain how they apply semantic web standards and tools to their scientific topic, from biology, agriculture, agro-ecology, genomics, environmental studies, to social sciences, citizen sciences.

Research papers presenting various aspects of semantic web technologies applied to biodiversity data, ranging from position papers to implemented systems descriptions and their evaluation have been selected. We have received 15 papers and selected 11 of them for the workshop, including 5 papers with long presentations and 6 papers with short presentations during the workshop. In addition, the workshop has offered two keynote presentations that are also mentioned hereafter.

We thanks the organization of the 10th ESWC 2013 for hosting the S4BioDiv workshop as a joint event and the Polytech'Montpellier engineering school for rooms and local arrangements in the beautiful city of Montpellier. Finally, we like to acknowledge our sponsors for the event, listed in the last page of the proceedings and thanks the scientific program committee for the reviewing of papers and discussion during the workshop presentations and panels.

The editors

Keynotes

Mark Wilkinson (Centro de Biotecnología y Genómica de Plantas UPM-INIA)

Web Science: A Distributed, Explicit, Transparent, Automated, Reusable, and Reproducible Experiments

Projections suggest that the delay between scientific discovery, and the dissemination and implementation of the knowledge embodied in that discovery, will soon vanish. At that point, all knowledge resulting from an investigation will be instantly interpreted and disseminated, influencing other researcher's experiments, and their results, immediately and transparently. This clearly requires that research results be of extremely high quality and reliability, and that research processes – from hypothesis to publication – become tightly integrated into the Web. Though the technologies necessary to achieve this kind of "Web Science" do not yet exist, our recently-published studies of automated in silico investigation demonstrate that we are enticingly close, and a path toward next-generation Web Science is now clear. The Web, to date, has only cosmetically changed the research process. Semantic Web Science, however, re-defines scientific methodology by fully integrating it with a global network of knowledge and expertise on the Semantic Web.

Olivier Rovellotti (Natural Solutions)

Semantic for Biodiversity: a user's perspective

In order to provide enlightened governance of our biodiversity heritage; it is crucial to gather and analyze as much biodiversity observational data as possible. Data gathering programs can be plotted on a scale of complexity and scope, from citizen science to professional environmental assessments. The data collected is so heterogeneous in quality, granularity and precision that it requires advanced data management techniques. Using semantic web technologies allows us to give various agents the correct tool to assist them in the entire process. Throughout our daily work in improving the data gathering, data aggregation and data visualization, we are able to give feedback on ours attempts at integrating semantic web technology in practical solutions.

Workshop articles

A Logical Model for Taxonomic Concepts for Expanding Knowledge using Linked Open Data

Rathachai Chawuthai[1], Hideaki Takeda[2], Vilas Wuwongse[3], and Utsugi Jinbo[4]

[1] Asian Institute of Technology, Prathumtani, Thailand
rathachai.c@gmail.com
[2] National Institute of Informatics, Tokyo, Japan
takeda@nii.ac.jp
[3] Thammasat University, Prathumtani, Thailand
wvilas@engr.tu.ac.th
[4] National Museum of Nature and Science, Tokyo, Japan
ujinbo@kahaku.go.jp

Abstract. The variety of classification systems and the new discovery of taxonomists lead to the diversity of biological information, especially taxon concepts. The association among taxon concepts across research institutes is very difficult to establish, because there is no single interpretation of the name of a taxon concept. Owing to this difficulty, further integration of more biological knowledge is very complicated when they deal with many sources of data or depending on different taxon concepts. This research aimes to develop a framework for linking some multiple related taxon concepts across research repositories, and preserving background knowledge of their changes. Therefore, we propose a logical model for taxon concepts in Resource Description Framework (RDF). Herewith, we implement a prototype to demonstrate the feasibility of our approach. It has been found that our model can publish taxon information as linked data and, hence, with additional benefits from Linked Open Data (LOD) cloud.

Keywords. Biological data, Biodiversity informatics, Logical model, Linked data, Ontology, Semantic web, Taxon concept

1 Introduction

More than 1.4 million species throughout the world have been truly described and classified with appropriate naming depended upon their characteristics; such as, morphological characters, living behaviors, DNA sequences, etc. [1-2]. Many taxonomists have dedicated themselves to study living organisms, research, and publish their knowledge for over hundred years. However, their researches have not been completely shared across all researchers around the world. In addition, there is no consensus on classification systems among taxonomists. In other words, taxonomists might have different perspectives to classify and name living organisms. As a consequence, a same species often be classified and named differently [2]. For example, *Papilio*

xuthus Linnaeus, 1767, Chinese Yellow Swallowtail Butterfly, has also been given several names by several taxonomists, such as *xuthulus* Bremer, 1861, *chinensis* Neuburger, 1900, *koxinga* Fruhstorfer, 1908, and *neoxuthus* Fruhstorfer, 1908.

The progress of taxonomic studies frequently causes redefinition of taxon concept, a circumscription of the taxon [2]. For instance, two genera of owls, *Nyctea* and *Bubo*, were merged into the latter genus *Bubo*. Following the change of genera, the scientific name of a snowy owl *Nyctea scandiaca* has been subsequently changed to *Bubo scandiacus* in order to satisfy the convention of scientific name [3]. Thus, a scientific name and a taxonomic concept become lacking of a single interpretation in biological [5-6]. Due to such change of taxon names, one sometimes misses information of this species under the name of the old scientific name when he or she searches information by the new scientific name.

Moreover, some details make researchers be confused when a taxon changes its concept without the change of its taxon name. For example, recently *Picoides tridactylus* (Three-toad Woodpecker) was split into two species, *P. tridactylus* (Eurasian Three-toad Woodpecker) and *P. dorsalis* (American Three-toad Woodpecker) [12]. Although these two species are disjointed, a part of information of *P. tridactylus*, especially recorded before the year 2003, might include details of *P. dorsalis*. One could obtain imprecise information when he or she simply searches information by the name *Picoides tridactylus*. Therefore, a mechanism that enables to link among taxon concepts in the precise context is necessary.

Recently, there was a research about managing the change in scientific conception. The work applied semantic web to develop a meta-ontology of a biological name (TaxMeOn). It provides metadata for representing and managing the temporal change of scientific name from a unit of taxon concept to another unit, and emphasized how the biological names publish [7]. However, the management of name change is not enough for researchers. The correct interpretation with temporal context of concepts and reasons of their changes becomes necessity as well.

The purpose of our research is to formulate a logical model for preserving background knowledge of the change of taxon concepts, and link some related concepts together. We introduced ontology for collecting the change of taxon concepts, cause and effect of the change; and linked data resulting from the change of concepts. We considered to enhance CKA [9] approach to capture the changes of taxon concepts, and their context. We also reused taxonomic terms from LODAC [8], employed SKOS[1] vocabularies to manage the relationship between concepts, and publicized data to Linked Open Data[2] (LOD) Cloud. Moreover, we performed an implementation to prove the feasibility of our proposed model.

To begin our approach, the background, the goal, and the related work have been already reviewed in this section. Next, Section 2, we will illustrate some technologies to develop our approach, and introduce the logical model in RDF. Section 3 will present prototype and discuss about its outcome. Lastly, Section 4 will draw conclusions and suggest some future improvements.

[1] Simple Knowledge Organization System: `http://www.w3.org/TR/skos-primer/`
[2] Linked Open Data: `http://linkeddata.org/`

2 The Proposed Logical Model

In this section, to achieve our objectives, we introduced a logical model for taxonomic concepts for expanding knowledge using LOD. Here, our model is expressed in ontology named Linked Taxonomic Knowledge (LTK) which was enhanced from several existing approaches.

Firstly, we studied how to classify the change of taxon concept; we found that they are two major categories: the change of name, and the change of classification [2,7,11]. A taxon name is sometimes changed for several reasons. For example, Hoare (2008) established the genus *Kendrickia* (ostracods). Then Kempf (2010) found that this genus was a primary junior homonym of *Kendrickia* Solem, 1985 (gastropods), and proposed *Dickhoarea* as the replacement name for *Kendrickia* Hoare, 2008. It results to the subsequent change of species names; for instance *Kendrickia asketos* had been changed into *Dickhoarea asketos* since Kampf (2010) has been published [2]. Apart from such name change, classifications also may be changed according to the progress of taxonomic researches. For example, the genus *Columba* (pigeons) has been split into five genera: *Patagioenas*, *Chloroenas*, *Lepidoenas*, *Oenoenas*, and *Columba* in the new narrow concept, and then some species of genus *Columba* have been assigned to one of these newly separated genera [12]. For instance, *Columba speciosa* changed to *Patagioenas speciosa* [12]. The analysis of the changes of taxon concept is described by Fig. 1.

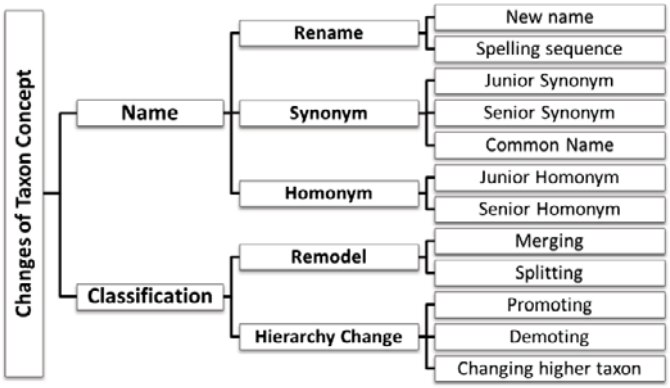

Fig. 1. The analysis of the changes of taxon concept

Secondly, we reviewed ideas in TaxMeOn, to describe concepts in taxonomic field linked to identifiers [7]. In general, when a concept's scope is changed, the changed concept needs to be recognized as new identifier. For instance, the genus *Bubo* before merging with *Nyctea* must not be the same identifier as *Bubo* after merging [2-3]. Thus, an identifier similar to those in TaxMeOn is required to our model. On the other hand, most attributes of the old *Bubo* can be copied to the new *Bubo* definitely, because, the old *Bubo* and the new *Bubo* may share many attributes.

Next, to publish data, we reviewed some standards that can be reused for our model. To model the entities of taxon concepts, we considered reusing some vocabularies from Linked Open Data for ACademia (LODAC), a project to publish a wide range of academic data including species information [8]. For example, a relationship between a species and a genus can be described as RDF using LODAC terms (species and genus are namespaces for species and genus in LODAC, respectively):

```
species:Nyctea_scandiaca  species:hasSuperTaxon     genus:Nyctea .
```

Another issue is to describe changes of concept and associated information on the change. There is an approach named Contextual Knowledge for Archives (CKA) Ontology. It offers a logical model developed using Flouris's theory for presenting the changes of conceptions, such as, merge, replace, and split. It also presents reasons behind the changes, changes of relationships such as the reclassification of a concept, and links between some relevant concepts. The CKA illustrates the change of concepts as dynamic RDF that contains fact and temporal aspect [9-10]. For instance, the following RDF expresses the splitting of a genus *Columba*.

```
ex:change2003 cka:interval          [tl:beginAtDateTime "2003"] ;
              cka:assure            ex:split1 .
ex:split1   rdf:type              ltk:TaxonSplitter ;
            cka:conceptBefore     genus:Columba ;
            cka:conceptAfter      genus:Patagioenas, genus:Chloroenas,
                                  genus:Lepidoenas,  genus:Oenoenas,
                                  genus:Columba_2003.
```

Further, the framework provides a technique to transform this dynamic RDF to static RDF with a given specific time point. For example, after year 2003, relationships among genus:Columba and its allies can be found as follows:

```
genus:Columba ltk:splitInto      genus:Patagioenas, genus:Chloroenas,
                                 genus:Lepidoenas,  genus:Oenoenas,
                                 genus:Columba_2003.
```

Technically, some operations from CKA framework can be extended to record the change of some concept's details, such as, color, size, organ, behavior, etc. It can be done by defining some new operations of change, and then binding the new operations with some related attributes. In addition, this model states one change as one unit. It offers association among related units of some changes by having some properties: cka:caused, and cka:effect to express reason and outcome of a change respectively. For example, Fig. 2 demonstrates the new name *Patagioenas speciosa* and its background. Consequently, we can find out the history of the name "*Patagioenas speciosa*". Then, we can use its background concept, such as the old name "*Columba speciosa*" to explore more information in the public LOD.

Fig. 2. Change of a taxon concept and its background

Lastly, to link data with LOD Cloud, our research proposed some useful operations that specify the change of taxon concepts, the changes of details of a taxon concept, the changes of relationships between taxon concepts, and the background of the change. All operations are defined by extending some vocabularies from the well-known ontology: Simple Knowledge Organization System (SKOS), and some properties from LODAC and CKA. Thus, the data from our framework can definitely be exchanged among other repositories. Example of some properties is shown in Table 1.

Table 1. Example properties from LTK which are derived from CKA, LODAC, and SKOS

Properties	rdfs:subPropertyOf
ltk:higherTaxon	cka:higherClass, skos:broaderTransitive, and species:hasSuperTaxon
ltk:replacedTo	cka:serialLinkTo, and skos:exactMatch
ltk:mergedInto	cka:serialLinkTo, and skos:relatedMatch
ltk:majorMergedInto	cka:serialLinkTo, and skos:closeMatch
ltk:synonym	skos:exactMatch

For example, the genus:Nyctea and genus:Bubo in old concepts have been merged into a new concept with the name *Bubo*. As stated previously, the genus *Bubo* in the new concept should be given a new identifier. In practice, we ended the year when it has been changed, so the new identifier of genus:Bubo may be genus:Bubo_1999. The property named ltk:mergedInto is defined to express a merge of two taxon concepts. The relationship between genus:Nyctea and genus:Bubo_1999 remains to be specified by the property ltk:mergedInto. On the other hand, another special property name ltk:majorMergedInto is introduced to demonstrate the very close relationship of two concepts, such as genus:Bubo and genus:Bubo_1999. As *Nyctea* was merged to *Bubo*, *Nyctea scandiaca*, the only member species of *Nyctea,* is transferred to *Bubo* and change the name to *Bubo scandiacus* [2-3]. In summary, these facts will be presented in RDF that satisfies the logical model of the CKA approach as follows:

```
ex:change1999    bibo:performer       pp:Wing, pp:Heidrich ;
                 bibo:issuer          pp:Richard ;
                 dcterms:source       pub:5224773;
                 cka:interval         [tl:beginAtDateTime "1999"] ;
                 cka:assure           ex:mg1, ex:rp1, ex:ac1 .
ex:mg1           rdf:type             ltk:TaxonMerger ;
                 cka:conceptBefore    genus:Bubo, genus:Nyctea ;
                 cka:conceptAfter     genus:Bubo_1999 .
ex:rp1           rdf:type             ltk:TaxonReplacement ;
                 cka:conceptBefore    species:Nyctea_scandiaca ;
                 cka:conceptAfter     species:Bubo_scandiacus .
ex:ac1           rdf:type             ltk:HigherTaxonAddition ;
                 cka:child            species:Bubo_scandiacus ;
                 cka:parent           species:Bubo_1999 .
ex:mg1           cka:cause            ex:rp1 .
ex:rp1           cka:detail           ex:ac1 .
```

After that, we apply some rules to transform dynamic RDF data to static form. For example, a rule that infers the merging operation of taxon concepts is expressed along these lines:

```
?change   rdf:type           ltk:TaxonMerger .
?change   cka:conceptBefore  ?before .
?change   cka:conceptAfter   ?after .
-------------------------------------------------
?before   ltk:mergedInto     ?after .
```

This rule and some others rules that infer each operation of change can convert the temporal RDF to be the following result.

```
genus:Nyctea              ltk:mergedInto          genus:Bubo_1999 .
genus:Bubo                ltk:majorMergedInto     genus:Bubo_1999 .
species:Bubo_scandiacus   ltk:higherTaxon    genus:Bubo_1999 .
species:Bubo_scandiacus   ltk:synonym        species:Nyctea_scandiaca .
genus:Nyctea                   cka:expired         "1999" .
genus:Bubo                     cka:expired         "1999" .
genus:Bubo_1999                cka:entered         "1999" .
species:Nyctea_scandiaca       cka:expired         "1999" .
species:Bubo_scandiacus        cka:entered         "1999" .
```

Therefore, clients can query these facts conveniently. For instance, if the users query some genera, which closely match (skos:closeMatch) genus:Nyctea; they will get genus:Bubo_1999. They sometimes query the data with species:hasSuperTaxon and get the result as same as ltk:higherTaxon. They can also find the present-day taxon concepts by inquiring some concepts which do not have a property named cka:expired. Moreover; the client can query more detail about a fact that includes the time when it changed, people who involved, reference documents, and triple data. For example, the replacement of species:Nyctea_scandiaca was caused by the merging between genus:Nyctea and genus:Bubo. In addition, the relationships of concepts can be presented by RDF statements, because the operation ltk:HigherTaxonAddition can establish the associations between concepts by producing some triples with having a property named ltk:higherTaxon. Our work offers some operations binding with properties; such as, dwc:scientificName[3], foaf:depiction[4], species:hasCommonName [8], etc. Thus, the consumers can query temporal information of taxon concepts along with specific time point.

3 Implementation and Discussion

After developing the LTK ontology, we verified the possibility and feasibility of it by implementing a prototype. The prototype is a web-based system that comprises three service layers: web interface, web services, and RDF data store. Firstly, the web interface allows a user to create the knowledge of taxon concepts in RDF. It also demonstrates the temporal context and link of taxon concepts. Further, it presents the reasons and details about changes of them. Secondly, the Java servlet service is made for

[3] Darwin Core Terms: http://rs.tdwg.org/dwc/terms/

[4] Friend of a Friend: http://xmlns.com/foaf/0.1/

managing and computing RDF data by using the performance of Jena[5] reasoning engine. Other clients can access data via this layer. Lastly, we used SESAME[6], a RDF store, to record data. Users can create data which come from some publications or books, and then the data is published to LOD cloud by providing SPARQL endpoint.

In Fig. 3, the left-side screen presents the context of the species:Nyctea_scandiaca (the figure displays as spc:Nyctea_scandiaca) and its linked taxon concepts, and the right-side screen shows information about the reason of changing this species. The web interface allows user to enter URI of concept and a specific time point in order to display the temporal context information as well.

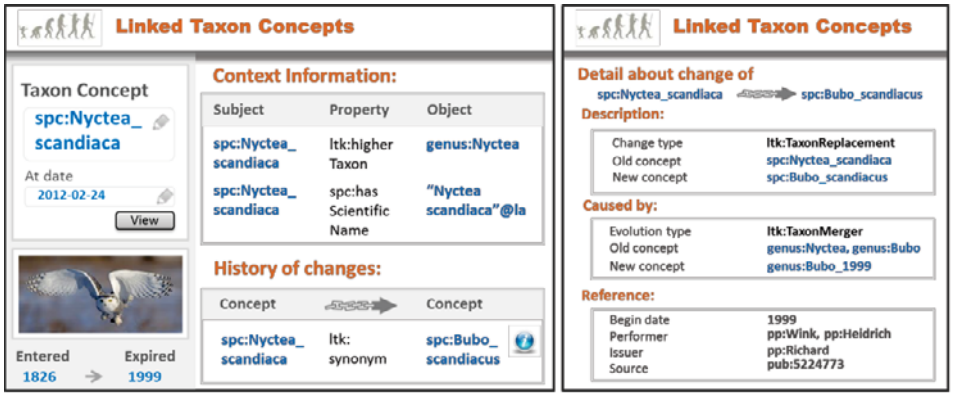

Fig. 3. Example screen of information about the concept species:Nyctea_scandiaca

As example RDF data in section 2, one change consists of many triples. When all changes are recorded, the triple store will manage over billion triples. Thus, it will consume a lot of resources when the service transforms the dynamic data to flat data for every request. However, most of all requests always ask for the present data. The prototype has to prepare current static data every time when each dynamic data is recorded. Then, the service can provide fast responses for the present information.

In summary, the prototype indicated that our approach is possible and feasible to make a real system. Moreover, other services can retrieve this data from LOD cloud.

4 Conclusions and Future work

Our paper presents a logical model and ontology for linking taxon concepts which comprises a series of changes, the diversity of taxonomic classifications, and the variety of naming. For the purpose of linking data, we have developed our model by employing ontology of contextual knowledge evolution together with some widely accepted ontology such as LODAC and SKOS. Therefore, our model can deal with both dynamic and static information represented in RDF and hence can trace the history of

[5] Apache Jena - reasoners and rule engines: `http://jena.apache.org/`
[6] SESAME – a framework for processing RDF data: `http://www.openrdf.org/`

the taxon concept. In addition, we have implemented a prototype which utilizes the proposed model in order to publish the taxonomic information to LOD cloud. As a consequence, other applications that need linked taxon concepts can readily connect to these data. Moreover, we have implemented a knowledge base using Jena's inference engine and SESAME's storage for computing data, and we have provided a web application to record and present the information. The result from our prototype demonstrates that our approach is feasible and suitable for the need of linked taxon concepts across different repositories and relationship backgrounds in order to discover broader knowledge of biology.

However, our approach gives priority to ontology rather than software application; hence the system requires much human effort to import a great number of data. For example, when a genus is split, some species under the genus have to move to new genera. In this case, taxonomists have to analyze and enter data by themselves. Thus, it should have some algorithms to improve the reclassification of some taxonomic ranks by their attributes. Moreover, in the future, when the number of data is over a billion, requesting historical data would be a great challenge because it requires the inference engine to process complex activities that consume very high computing capability. Future research might be focusing on how to improve the computing resources or methodologies for caching time-series of taxonomic data.

References

1. Darwin, C., Peckham, M.: On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. Penn Press, Philadelphia (1959)
2. Winston, J.E.: Describing Species: Practical Taxonomic Procedure for Biologists. Columbia University Press, New York (1999)
3. Wink, M., Heidrich, P.: Molecular evolution and systematics of the owls (Strigiformes). In: A Guide to Owls of the World. Yale University Press, Yale (1999)
4. International Commission on Zoological Nomenclature: International Code of Zoological Nomenclature. The International Trust for Zoological Nomenclature, London (1999)
5. Mallet, J.: Species, Concept of. In: Encyclopedia of Biodiversity. Elsevier, Oxford (2007)
6. Ytow, N., Morse, D., Roberts, D.: Nomencurator: a nomenclatural history model to handle multiple taxonomic views. In: Biological Journal of Linnean Society, pp. 81-98 (2001).
7. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological names and taxonomies on the semantic web: managing the change in scientific conception. In: ESWC 2011. LNCS, vol. 6644, pp. 255-269. Springer, Heidelberg (2011)
8. Linked Open Data for Academia, `http://lod.ac/`
9. Chawuthai, R., Wuwongse, V., Takeda, H.: A Formal Approach to the Modelling of Digital Archives. In: ICADL 2012. LNCS, vol. 7634, pp. 179-188. Springer, Heidelberg (2012)
10. Flouris, G., Meghini, C.: Terminology and Wish List for a Formal Theory of Preservation. In: PV 2007. Proceedings, DLR, Munich (2007)
11. Franz, N., Peet, R.: Towards a language for mapping relationships among taxonomic concepts. In: Systematics and Biodiversity, vol. 7, iss. 1, pp. 5-20 (2009)
12. Banks, R.C., Cicero, C., et al.: Forty-fourth supplement to the American Ornithologists' Union check-list of North American birds. In: The Auk, vol. 120, pp. 923–931 (2003)

Publishing and Using Plant Names as an Ontology Service

Jouni Tuominen, Nina Laurenne, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Aalto University School of Science, Dept. of Media Technology, and
University of Helsinki, Dept. of Computer Science
http://www.seco.tkk.fi, firstname.lastname@aalto.fi

Abstract. Animals and plants are referred to using scientific or common names depending on the expertise of an audience or a source of data. The names change in time and therefore their usage as identifiers as such is problematic. We present a solution for managing and using plant names as an ontology. The ontology is based on the TaxMeOn meta-ontology for biological names. In order to refer to organisms unambiguously and publish information as Linked Data on the web, the names are given URIs. The ontology is developed collaboratively and it supports the approval process and temporal tracking of the common names. We introduce an ontology service of plant names for end-users and provide user interfaces and APIs for integrating the ontology into applications.

1 Introduction

The scientific names of plants and animals have a major role when indexing, querying, and integrating information about species. Biologists use scientific names although the vast majority of people use the common name equivalents. Contrary to common belief, neither the scientific nor common names identify organisms unambiguously as one name may point to multiple species and one species may have multiple names.

New research results change the name combination of the scientific names because taxa are constantly split and lumped. For example, if a species is changed into another genus, the name combination changes accordingly. Approximately 25,000 new species descriptions are published in thousands of journals annually [6] which makes it hard for researchers to keep up-to-date the biodiversity of the nature. Not all organisms need a common name but still there is huge work to be done in developing the vernacular nomenclature and in terms of established names, the dialect expressions remarkably expand the spectrum of the biological names.

The international commissions of the nomenclatures (IBC, ICZN) specify the rules how the scientific names should be used in various taxonomic treatments. The nomenclatures of plants and animals are independent of each other and the rules are applied only to the scientific names. The common names are not

regulated but they also change in time because there is often a need to update the common names at intervals. The changing nature of the names poses challenges for their management [5, 10, 13].

The diversity of the names causes problems when combining data from heterogeneous sources, e.g., observational records, literature and museum collections [11, 9]. The data cannot be easily integrated if a taxon is referred to using multiple names and vice versa the existence of homonyms (the same name refers to multiple taxa) causes errors when merging the data.

Comprehensive reference lists and catalogues of the names have been proposed as a solution to facilitate the access to the names [1, 10]. However, this is not enough because the biological names ought to be machine-processable in order to refer to them unambiguously and semantically enrich the biological contents. Ontologies remarkably increase the re-use and utilization of the available resources which minimizes the amount of manual work when harmonizing data.

We present an ontology model for managing the common names of organisms and linking them to the scientific names. The model supports temporal tracking of name changes and an approval process of the common names. The model is used for maintaining and publishing plant names in Finnish as an ontology. The ontology is published as Linked Open Data [3] and can be used as an ontology service.

2 Ontology Model

TaxMeOn[1] [14] is an RDF-based meta-ontology for modeling and managing biological names and classifications. TaxMeOn introduces classes and properties for expressing biological names as ontologies. The model consists of three parts according to the level of taxonomic details, which are common names, species checklists, and detailed taxonomic information respectively. In this paper, the focus is on the common names although many of the classes and properties are common to all three parts. The simplified structure of the model is presented in Fig. 1, where the core classes are *Scientific name*, *Common name* and their statuses. The status of the *Scientific name* indicates if a name is an accepted or a synonymized one, etc. The synonyms are linked to an accepted name. The hierarchical structure is constructed setting relations between the *Scientific names*.

The *Common names* (in one or more languages) that refer to the same taxon are connected through a *Scientific name*. The model also allows mapping the scientific names to each other based on the underlying taxonomic concepts (congruence, overlap, part-of, general association). A taxonomic rank expresses the hierarchical level in a classification (e.g., a species, a genus) and it is specified for every scientific name. The taxonomic ranks are presented as a separate vocabulary which contains 61 ranks, of which 60 are obtained from TDWG Taxon Rank LSID Ontology[2]. In order to avoid the complex details of the botanical and

[1] `http://schema.onki.fi/taxmeon/`

[2] `http://rs.tdwg.org/ontology/voc/TaxonRank`

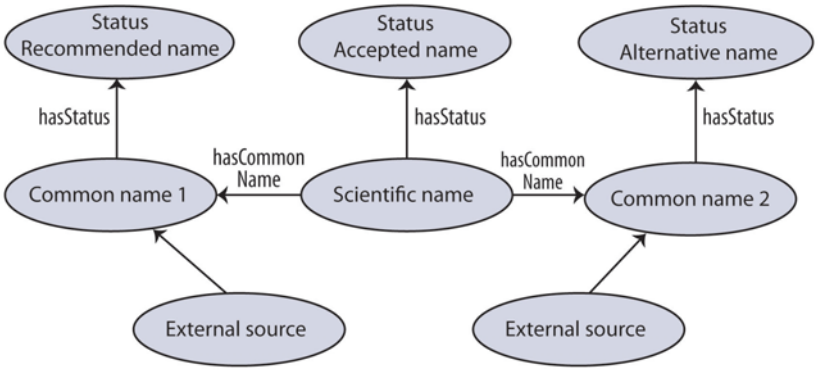

Fig. 1. The ontology model of the common names of organisms. The ellipses represent classes and the arrows depict relations between the classes.

horticultural nomenclatures, the species level and the taxonomic levels below it are treated as one unit.

The approval process of the common names is the following: first, a new name is proposed; then the name becomes accepted; and finally, the name may become an alternative, if there is a new accepted name for the same plant. The model allows the maintainers to propose a new name which then can be commented by the other maintainers until the name finally gets accepted, rejected or synonymized. The temporal management of the names is based on time stamps which are given to the statuses of the names in the approval process. If a name is given a new status, the old status is not removed from the system. This makes it possible to track the chain of changes of the names and to see the period of time period when a particular name was accepted.

3 Managing Plant Names as an Ontology

We applied the TaxMeOn ontology model to a database of the Finnish names of plants maintained by the Finnish Biology Society Vanamo[3]. The original database contained nearly 26,000 plant names in Finnish in a single classification. The taxa were divided into three taxonomic levels (a species, a genus and a family) but it is possible to specify more taxonomic levels in the current ontology.

The database of the plant names was converted into RDF format based on the TaxMeOn ontology model. The ontology is managed in the metadata editor SAHA[4] [7] by the Vanamo association. Currently, the ontology contains 21,797 species, and the number of updates exceeds one thousand names yearly. The

[3] `http://www.vanamo.fi`
[4] `http://www.seco.tkk.fi/services/saha/`

utilization of the ontology facilitates the management of the names because the approval process is integrated into the ontology.

The association has an active role in developing new Finnish names for plants and the public availability of the ontology releases voluntary based work for more relevant activities than responding to various queries by journalists, translators etc. The development of the new names is based on the needs, therefore the coverage of the taxa is not systematically or geographically restricted into any particular plant group or a region.

The browser-based SAHA editor allows collaborative editing of the ontology, providing the simultaneous access of multiple users and a chat functionality. The TaxMeOn model has been extended to support the management of the ontology in SAHA, by adding a property indicating the current status of the processing of a proposed common name. If a new name is suggested for a species, a maintainer can add it into the ontology and mark it as "in process". The proposed but not yet processed names can be found easily at later stages of the process.

4 Using Plant Names as an Ontology Service

The ontology is published as Linked Open Data in the Finnish Ontology Library Service ONKI[5] [15], as part of the Finnish semantic web infrastructure project FinnONTO[6] [4]. The ONKI service provides user interfaces and APIs for accessing and using the plant names in applications. For example, end-users can browse and search the ontology to find a common name for a taxon that they know only by the scientific name. The ONKI selector widget can be integrated into legacy CMS systems to provide an autocomplete and URI fetching features to support the annotation of plant related information.

One of the advantages of the ontology service is that the end-users can now access the ontology themselves. Users are directed to the ONKI service via search engines, and they have adopted the service by extending Wikipedia articles about plant species with links to Finnish plant names in ONKI. End-users actively send feedback, comments and corrections to the maintainers, which help them to improve the quality of the content.

The ontology is also accessible as a SPARQL endpoint. An example query below shows how the accepted Finnish common names of species (and taxa below it) that belong to a genus "*Quercus*" (oak) can be retrieved:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX taxmeon: <http://www.yso.fi/onto/taxmeon/>
PREFIX taxonomic-ranks: <http://www.yso.fi/onto/taxonomic-ranks/>

SELECT ?vernacularName WHERE {
  ?species taxmeon:isPartOfHigherTaxon ?genus .
  ?genus rdf:type taxonomic-ranks:Genus .
  ?genus rdfs:label "Quercus"^^xsd:string .
```

[5] http://onki.fi/en/browser/overview/kassu

[6] http://www.seco.tkk.fi/projects/finnonto/

```
  ?species taxmeon:hasVernacularName ?vernacularNameRes .
  ?vernacularNameRes taxmeon:hasVernacularNameStatus ?status .
  ?status rdf:type taxmeon:AcceptedVernacularName .
  ?vernacularNameRes rdfs:label ?vernacularName .
  FILTER langMatches(lang(?vernacularName), "fi")
}
```

The result of the query is a list of the Finnish names of oak species, such as the sessile oak and white oak. The query demonstrates the use of the ontology for cross-language query expansion.

Currently, the plant name ontology is used by several cultural museums and libraries for annotating their collections. The ontology is also applied as a use case in the EU FP project ENVIROFI[7] which focuses on the environmental usage area of the Future Internet. The ontology is used in the project as a conceptual hub for referring to the plants in the observational data on biodiversity. The ontology has been extended with the English and German names of plants used in the project pilots (these names are not available in the ONKI ontology service).

5 Discussion

5.1 Related Work

The importance of persistent identifiers for organism names and solutions for managing them on the semantic web have been discussed by several workers. Page [8] presented how taxon names are modeled as semantic metadata in RDF form. Taxon names are identified with using Life Science Identifiers (LSID) and the names are connected using taxonomic relations. Taxon names that are obtained from various data sources and which refer to the same taxon are mapped using the *owl:sameAs* relation. Schulz et al. [12] presented the first ontology model of biological taxa and its application to physical individuals. The model is based on a single unchangeable classification. Franz and Thau [2] evaluated the limitations of applying ontologies to the scientific names and concluded that ontologies should focus either on a nomenclatural point of view or on strategies for aligning multiple taxonomies.

The Darwin Core (DwC)[8] is a metadata schema developed for taxon occurrence data by the TDWG (Biodiversity Information Standards). The goal of DwC is to standardize the form of how biological information is presented. However, it lacks the semantic aspect and when it comes to the names, the scope of DwC is quite general.

Taxonconcept.org[9] provides Linked Open Data identifiers for species concepts and links data from different sources. All the names of species are expressed using literals. Also, the machine-processability is weakened by the usage of literal values for expressing the hierarchies. The data contains scientific and common names, and taxonomic statuses.

[7] http://www.envirofi.eu
[8] http://www.tdwg.org/standards/450/
[9] http://www.taxonconcept.org

Many existing databases aim to be comprehensive online catalogues that aggregate individual species checklists, such as the Catalogue of Life (CoL)[10] and The International Plant Names Index (IPNI)[11]. The IPNI database contains only scientific names, but the Catalogue of Life also includes their taxonomic statuses and common names. They both provide the names in a machine-processable form, as RDF conforming to the TDWG Taxonomic Concept Transfer Schema (TCS)[12] using LSIDs as identifiers of the names [5]. In the Catalogue of Life the requirement to use a separate LSID resolver for fetching metadata about an LSID prevents the Linked Data compatibility of the dataset. The IPNI database provides an LSID proxy that allows Linked Data compatibility. In the IPNI database, the hierarchy is not expressed explicitly in the RDF (e.g., the genus of a species is shown only in the binomial name literal).

There are several other plant name databases available on the web, e.g., the Royal Horticultural Society Horticultural Database[13], The Plant List[14] and the Euro+Med PlantBase[15]. Most available resources contain the scientific names, but in few, the common names are included. Common to these systems is that they are intended for human usage, and they are not available in a machine-processable form with unique name identifiers.

5.2 Contributions and Future Work

Most of the related work concentrate on the scientific names, but our focus is on the common names. The common names expand the cross-domain use of the ontology because they are in wider spectrum of use than the scientific ones. The ontology is available in machine-processable RDF format, with explicit semantics, e.g., the hierarchical relations are set between the plant URIs, and the statuses of names are supported. The TaxMeOn model provides a solution for managing the approval process of common names, supporting the temporal tracking of the name changes via statuses and their time stamps. The model connects together different names of a taxon facilitating data integration and information retrieval in cases where data is combined from heterogeneous sources.

We have also demonstrated the complete workflow from a collaborative development of an ontology to publishing it as Linked Open Data and as an ontology service which makes it accessible to the general public. The plant name ontology helps harmonizing the terminology which in turn enhances communication between various users. Application developers can utilize the ontology by using the plant name URIs for unambiguous referencing to plants species.

Currently, hybrid taxa are modeled in the ontology in the same way as the ordinary species. An idea for the future development is to extend the model to

[10] `http://www.catalogueoflife.org`
[11] `http://www.ipni.org`
[12] `http://www.tdwg.org/standards/117/`
[13] `http://apps.rhs.org.uk/horticulturaldatabase`
[14] `http://www.theplantlist.org`
[15] `http://www.emplantbase.org`

support the representation of hybrid names at a deeper level. Another area for development is to link the scientific names of plants to their author URIs in DBpedia, connecting the ontology to the Linked Data Cloud (LOD).

Ontologies are a bridge between experts and ordinary people in communication and popularizing science. Additionally, the Linked Data approach provides a way how to easily extend an ontology with additional information which in turn increases the information value of contents.

Acknowledgments This work is part of the National Semantic Web Ontology project in Finland FinnONTO (2003-2012), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 public organizations and companies, and the EU FP project The Environmental Observation Web and its Service Applications within the Future Internet (ENVIROFI). We thank Leo Junikka and Arto Kurtto for their collaboration.

References

1. Dengler, J., Berendsohn, W.G., Bergmeier, E., Chytrý, M., Danihelka, J., Jansen, F., Kusber, W.H., Landucci, F., Müller, A., Panfili, E., Schaminée, J.H.J., Venanzoni, R., von Raab-Straube, E.: The need for and the requirements of EuroSL, an electronic taxonomic reference list of all european plants. Biodiversity & Ecology 4, 15–24 (2012)
2. Franz, N., Thau, D.: Biological taxonomy and ontology development: scope and limitations. Biodiversity Informatics 7, 45–66 (2010)
3. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1–136, Morgan & Claypool (2011)
4. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In: Proceedings of the ESWC 2008, Tenerife, Spain. Springer–Verlag (2008)
5. Jones, A.C., White, R.J., Orme, E.R.: Identifying and relating biological concepts in the Catalogue of Life. Biomedical Semantics 2(7) (2011)
6. Knapp, S., Polaszek, A., Watson, M.: Spreading the word. Nature 446, 261–262 (2007)
7. Kurki, J., Hyvönen, E.: Collaborative metadata editor integrated with ontology services and faceted portals. In: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece. CEUR Workshop Proceedings, `http://ceur-ws.org` (2010)
8. Page, R.: Taxonomic names, metadata, and the semantic web. Biodiversity Informatics 3, 1–15 (2006)
9. Page, R.D.M.: Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics 9(5), 345–354 (2008)
10. Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L., Remsen, D.P.: Names are key to the big new biology. Trends in Ecology & Evolution 25(12), 686–691 (2010)
11. Sarkar, I.N.: Biodiversity informatics: organizing and linking information across the spectrum of life. Briefings in Bioinformatics 8(5), 347–357 (2007)

12. Schulz, S., Stenzhorn, H., Boeker, M.: The ontology of biological taxa. Bioinformatics 24(13), 313–321 (2008)
13. Segers, H., de Smet, W.H., Fischer, C., Fontaneto, D., Michaloudi, E., Wallace, R.L., Jersabek, C.D.: Towards a list of available names in zoology, partim phylum rotifera. Zootaxa 3179, 61–68 (2012)
14. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological names and taxonomies on the semantic web – managing the change in scientific conception. In: Proceedings of the ESWC 2011, Heraklion, Greece. pp. 255–269. Springer–Verlag (2011)
15. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The Finnish ontology library service ONKI. In: Proceedings of the ESWC 2009, Heraklion, Greece. pp. 781–795. Springer–Verlag (2009)

A faceted search system for facilitating discovery-driven scientific activities: a use case from functional ecology

Marie-Angélique Laporte[1], Eric Garnier[2], and Isabelle Mougenot[1]

[1] UMR 228 ESPACE-DEV, Maison de la Télédétection 34093 Montpellier, France,
firstname.lastname@ird.fr
[2] Centre d'Ecologie Fonctionnelle et Evolutive (UMR 5175), 1919 Route de Mende, 34293 Montpellier Cedex 5, France,
firstname.lastname@cefe.cnrs.fr

Abstract. To address biodiversity issues in ecology and assess the consequences of ecosystem changes, large quantities of long-term observational data from multiple data sets need to be integrated and characterized in a unified way. During these last decades, functional trait-based approaches have shown great potential to facilitate the understanding and the prediction of ecosystem changes. To promote data exchange, portability and to drive higher communiction between systems, scientific communities are required to acquire data standards. Semantic web (or web of data) provides a realistic solution for these exact requirements. Consequently semantic web allows for creative approaches and offers opportunities to scientists to gain new insight from experimental data. A first step to this goal is to standardize meaningful and precise terms that are interlinked through a dedicated thesaurus that covers the plant functional diversity domain. Therefore this vocabulary can serve as stable reference resources for integration purposes, specifically when published in RDF language and available as linked data on the web. This manuscript presents a web infrastructure, named Thesauform, that fully exploits the key principles of the web of data and its common open data structures in order to guide the plant functional diversity community of experts to build collectively, manage, visualize and query a SKOS thesaurus. A thesaurus dedicated to plant functional traits is used to demonstrate the potential of the approach. Indeed, the thesaurus, built using the Thesauform tool, is used to semantically annotate heterogeneous data sources, such as the TRY database or the Plant Ontology. Then, a faceted search system, based on SKOS collections, enabling thesaurus browsing according to each end-users requirements is expected to greatly enhance the data discovery in the context of biodiversity studies.

Keywords: Tool, Faceted Search, Thesaurus, Semantic annotation, Functional diversity, Web of Data, Plant Trait, Controlled vocabulary, Interoperability, SKOS

1 Introduction

Resolution of key biodiversity issues goes through continued exchanges and cooperation between related domains, such as ecology, taxonomy, genomic, climatology, soil sciences, etc [1], [2], [3]. To address biodiversity issues, it is now widely accepted that a functional approach has strong potential. Indeed, biological traits of organisms have great capabilities to promote a better understanding and to predict global change consequences on the functioning of ecosystems and the services they provide to human societies [4], [5], [6], [7]. A functional trait is defined as: "any morphological, physiological or phenological feature measurable at the individual level, from the cell to the whole-organism level, without reference to the environment or any other level of organization" [8].

Over the last decades, trait-based research has generated huge volumes of data, within multiple contexts of observations and experiments [9]. Considering this, data can be acquired via specific studies and are influenced by peculiar goals. Additionally, these data sets can also be obtained via very different study contexts and are often described in highly specialized terms. Numerous traits can be measured, for instance, on plants [9], [10], [11], [12]. However, data representation and storage do not constitute a major challenge. This is why data generated by functional ecology are only minimally reused or shared within the community, or over communities, mainly due to data heterogeneity. Given these limitations, open web standards and the generation of open web standards for functional ecology would advance the integration of heterogeneous content, with the primary objective of the emergence of new knowledge.

Our primary concern, which focuses on access, sharing and dissemination of information within a community of experts, is oriented towards the semantic web. The web of data [13], [14] provides the concepts, methods and tools, which allow a gradual slide from a web that mostly supports sharing documents to a web that focuses on the sharing of data to ensure their joint and concerted use by software agents. The web of data is primarily based on the key principles of metadata and controlled vocabularies or even ontologies, which should be considered complementary. Thesaurus, which is a type of controlled vocabularies, bypass ambiguity issues in natural language, in order to control and to clarify the access and exchange of information and to facilitate communication. Consequently a thesaurus reflects deliberate choices of communities relatively to the key terms in their expertise field. SKOS (Simple Knowledge Organization System) [15] provides a common format to manage thesaurus adequately. The need for the simultaneous use of multiple vocabularies being increasing in a context of biodiversity studies, SKOS offers not only the mean to build and to publish a thesaurus on the web, but also to anticipate the establishment of cross-references between thesauri. Accordingly, each thesaurus can be considered as a publicly available relevant resource on the web and can be enriched via meaningful navigation between thesauri, when properly described in an adequate format.

In this paper, we present a complete system dedicated to the ecological community allowing it to create, manage, visualize and query a SKOS thesaurus. The final purpose of the thesaurus is to facilitate the integration and the navi-

gation of the information available in multiple data sources. Our previous work focused on how metadata could be exploited during the collaborative building of a thesaurus, through edition and extension mechanisms using the Thesauform tool. The functional plant trait thesaurus (TOP thesaurus, for Trait Of Plant Thesaurus) was built using the Thesauform tool. In this paper, our goal is now to demonstrate the full capabilities of the TOP thesaurus. First, the TOP thesaurus is used to establish mappings between TOP concepts and other data sources, as for instance the TRY database [9] and the Plant Ontology (PO) [16], [17], in a vision of open data sources, in order to both interconnect available information, and semantically annotate data organized into these data sources. Secondly, the TOP thesaurus is exploited through a faceted search engine that reflects end-users interests and preferences, to facilitate the appropriation of the TOP thesaurus by end-users. The facets then act as access points on the interrelated data sources in guiding their navigation. The TOP thesaurus then fully plays its role by supporting the functional plant trait community to manage existing and future datasets and to interconnect them with data from other relevant domains.

This article is organized as follow:

- Section 2 introduces the approach driven with the Thesauform tool to build a functional plant trait thesaurus as a collaborative product. Once the thesaurus has been built, it serves as stable reference resources for integration purposes and it is used to semantically annotate heterogeneous data sources, such as the TRY database or the Plant Ontology.

- Section 3 explains how faceted search enhances the information retrieval. This section gives an overview of the technologies used to query a SKOS thesaurus using end-user preferences.

- Section 4 consists of the implementation of our approach. This section presents the key features of the TOP thesaurus-browsing interface based on faceted search and how this interface is used by functional ecology expert to find relevant information about functional plant traits.

- Finally, section 5 summarizes and discusses the strengths of our approach and refers to future work.

2 Developing a collective thesaurus: the example of the TOP thesaurus

In order to build a collective thesaurus, our recent work focused on the development of a tool, named Thesauform, dedicated to assist domain experts in their task. The Thesauform tool fully relies on semantic web standards, while providing a flexible and user-friendly environment for domain experts [18]. The process of thesaurus co-construction was divided into two phases: (i) an edition phase, during which experts can perform a number of actions in relation to the construction of the vocabulary (addition/deletion of terms and concepts, change of definitions, addition of a commentary, etc.), and (ii) a validation phase, where experts can validate or invalidate the results of the activities completed

during the previous edition phase through a voting procedure. The functional plant trait community has used the Thesauform tool to describe the different functional plant traits in use in the domain.

A part of the TOP thesaurus, based upon the Thesauform, is shown on Table 1. Twenty different experts from the functional plant trait community collaboratively developed the TOP thesaurus. Currently the TOP thesaurus is composed of about 1200 terms regrouped into approximately 1000 concepts. The TOP thesaurus can be used as a bibliographic resource about plant traits information, since it is available as a web resource. For each trait concept, a preferred term, a definition associated to a bibliographic reference and a broader term are provided. In some cases, synonyms (alternative terms), abbreviation, related terms and narrower terms are also specified, as well as a preferred unit [3] . For instance, the widely used trait "Specific Leaf Area", also known under the abbreviation SLA, is defined as "the one sided area of a fresh leaf divided by its oven-dry mass" in Cornelissen et al. 2003, and its measurement unit is expressed in meter squared by kilogram of dry mass (m2kg-1[DM]). In the thesaurus, this trait is linked to different other traits. Indeed, it falls under the broader concept of Morphology and it is related to the Leaf Blade Thickness and the Leaf Mass per Area concepts.

The TOP thesaurus serves as a stable reference resource by organizing traits and their information. It extends beyond the users needs by linking information about traits to different available data sources with the purpose of both enriching and facilitating data interpretation, which requires information from different domains. Consequently, TOP thesaurus concepts have been linked to two different data sources, the TRY database and the Plant Ontology (PO). A real advantage of SKOS is to provide properties dedicated to the establishment of cross-references between thesauri. The mapping apporoaches, on one hand between the TOP thesaurus and the TRY database and on the other hand the TOP thesaurus and PO, rely on the exactMatch and relatedMatch SKOS properties.

The advantage of linking TOP thesaurus concepts to TRY, the biggest functional plant traits database (about 800 traits are measured in TRY on more than 60000 different plant species), is double. First, the mapping TOP/TRY allows TOP thesaurus to unify the access to TRY data, managing the heterogeneity terms used to describe TRY data. Secondly, such a mapping will show the TRY observation number and the geo-referenced observation number for each mapped trait, or the number of different species, on which a given trait has been measured. This information can be useful to account for both the community interest for a given trait or the data available on a trait.

[3] The SKOS vocabulary has been expanded to add this information to the TOP thesaurus, considering the importance of the measurement units for trait data interpretation and quality

Table 1. Modified from Laporte et al. 2012. A subset of the traits present in the TOP thesaurus, together with their information attached (definition and associated reference, broader term (if any), narrower term (if any), preferred unit).

Trait Prefered Label	**Définition**	**Reference of the defintion**	**Synonym Alternative Label**	**Abbr**	**Related Term**	**Broader term (BT) / Narrower term (NT)**	**Preferred unit**
Trait	Any morphological, physiological or phenological feature measurable at the individual level, from the cell to the whole-organism level	Violle et al., 2007				NT:Chemical composition, NT:Optical property, NT:Size, NT:Structure, NT:Time related	
Specific leaf area	The one sided area of a fresh leaf divided by its oven-dry mass	Cornelissen et al., 2003		SLA	Leaf blade thickness Leaf mass per area	BT:Morphology	m2kg-1[DM]
Leaf lifespan	The time period during which an individual leaf or part of a leaf is alive and physiologically active	Cornelissen et al., 2003	Leaf longevity			BT:Life cycle	months
Specific root length	The ratio of root length to root mass	Cornelissen et al., 2003		SRL		BT:Morphology	m.g-1
Plant height observed	The shortest distance between the upper boundary of the main photosynthetic tissues on a plant and the ground level	Cornelissen et al., 2003	Reproductive plant height		Canopy height	BT:Height, NT:Phanerophytes, NT:Chamaephytes, NT:Hemicryptophytes, NT:Cryptophytes, NT:Geophytes, NT:Therophytes, NT:Helophytes, NT:Hydrophytes	m

Trait Prefered Label	Définition	Reference of the defintion	Synonym Alternative Label	Abbr	Related Term	Broader term (BT) / Narrower term (NT)	Preferred unit
Leaf mass per area	The oven dry mass of a leaf divided by its one-sided area	Cornelissen et al., 2003	Leaf specific mass	LMA	Specific leaf area	BT:Morphology	g[DM]m-2
Leaf phenology	The timing of foliage of the whole canopy	Cornelissen et al., 2003				BT:Life cycle	unitless
Leaf photo-synthetic rate	The photosynthetic rate per unit leaf mass at measurement temperature	Hendry and Grime, 1993		JCO2		BT:Leaf photosynthetic rate	nmoles /cm2/sec
Wood density	The oven-dry mass of a section of a plant's main stem divided by the volume of the same section when still fresh	Cornelissen et al., 2003	Stem specific density			BT:Morphology	kg[DM]dm-3
Bark thickness	Tickness of the part of the stem that is external to the wood including vascular cambium	Cornelissen et al., 2003				BT:Anatomy	mm
Seed dry mass	The air-dried mass of a seed	Cornelissen et al., 2003	Diaspore mass	SM		BT:Seed mass	mg
Seed shape	The variance of the three dimensions length width and height dividing each dimension by length so length is unity	Thompson et al., 1993				BT:Morphology	unitless
Relative growth rate	The per unit rate if growth of a plant or plant part.	Evans, 1972		RGR		BT:Growth rate, NT:Whole plant relative growth rate, NT:Stem relative growth rate, NT:Root relative growth rate, NT:Leaf relative growth rate, NT:Shoot relative growth rate	g/g/d

PO, which is a controlled vocabulary describing plant entities, is of great interest for plant traits, since plant traits are measured on plant tissues or organs. The mapping established between TOP thesaurus concepts and PO concepts allow assigning a reference for the plant entities cited in most trait definitions. Moreover, such a mapping approach will be highly beneficial to link data used in ecology or agronomy to data used in genomics. In fact, PO is mainly used by this latter field and provides the opportunity to serve as a first unifying component between the ecological and the genomic world, both of high interest in biodiversity studies.

The TOP thesaurus fulfills its initial role to provide standard vocabulary available to the functional ecology community, and extends beyond the basic needs to ease information retrieval. In this context, a system considering end-user points of view has been developed and offers a faceted search engine.

3 Information retrieval: Faceted search

Information retrieval using free text search is confronted with limitations in terms of accuracy of the result [1], [19]. The use of controlled term and concepts coming from a thesaurus would enhance data queries [3]. Classic semantic search engines based on controlled terms have been widely used to query data in life science fields. Bioportal[4] [20] is a web portal providing the interrogation of multiple ontologies or controlled vocabularies based on controlled terms. This kind of search mechanism suffers from limitations, since it can be difficult for an unexperienced end-user to find relevant controlled terms. In fact, with classic semantic search engines, most of the time controlled terms are displayed using auto-completed search fields. To overcome this limitation, faceted search engines are an interesting solution as they facilitate the thesaurus appropriation by the end-users. In such search engines, the results are filtered using relevant parameters or categories, each category reflecting the need of users in the thesaurus navigation environment.

On the MUMIA [5] web site, faceted search (also called faceted navigation or faceted browsing) is defined as: "a technique for accessing a collection of information, allowing users to explore by filtering available information. A faceted classification system allows the assignment of multiple classifications to an object, enabling the classifications to be ordered in multiple ways, rather than in a single, pre-determined, taxonomic order". Each facet typically corresponds to the common features shared by a set of objects. Faceted searches are commonly used by e-commerce websites to filter the available products based on the parameters most important for the user choice.

Faceted search systems can be applied to SKOS thesaurus. SKOS good practices describe how to represent such a system in a SKOS compliant way [21]. Facets are closely linked to both thesaurus information visualization and thesaurus information restitution, but not to thesaurus structure or to the infor-

[4] http://bioportal.bioontology.org/
[5] http://www.mumia-network.eu/index.php/working-groups/wg4

mation it carries. In thesaurus or in any other controlled vocabulary, concepts can be assembled into semantically meaningful groups, corresponding to facets. Consequently, facets can be defined as skos:collection [21], [22], [23], gathering concepts with common features. For instance, the functional plant trait concept Specific Leaf Area (2) can be grouped with the concepts Leaf Phenology or Leaf Lifespan, because these three concepts share the common feature of being measured on the same plant part, the leaf. But Specific Leaf Area may also be classified with the Xylem Area concept, because these two measurements refer to a size measurement, the area. The categories plant organ and measurement type can then be consider as two access points to query thesaurus. Each user can choose, which access point to use according to own preference.

Faceted search system is so of prior interest to assist users in their information retrieval. Developing such a system based on facets allows taking users interest into account and then to guide dataset consultation. Having data sources semantically annotated with TOP thesaurus concepts can benefit from faceted search engines as well, because thesaurus facets are used as an access point to disseminate information from heterogeneous data sources.

4 Results: approach implementation, user interface

TOP thesaurus trait information will be mainly accessed by experts from the ecology domain. Considering this, we based our work on an user-friendly and easy to use interface, to assist experts in their access and retrieval of pertinent information. In this section we present the key features of our system[6].

4.1 Semantic search engine

Search is a crucial feature for focused information retrieval. We propose two types of semantic search approaches to access functional plant trait information (cf. 1). First, a classic semantic search engine is available and allows finding traits with controlled trait terms from the TOP thesaurus through an auto-completed field search and a navigation tree. A unique aspect of our work is the implementation of a faceted search engine based on skos collections. This enhanced the semantic search of trait by providing the opportunity to the users to choose his own filters. In Figure 1, end-user selected categories from the available facets (the selected categories are colored in green). The result of such a selection is dynamically updated in the result part.

4.2 External data sources mapping

To address to need of biodiversity studies and to enhance the cooperation and the sharing of heterogeneous data inside the functional plant trait community

[6] For the features concerning the collaborative thesaurus building using the Thesauform tool, please refer to Laporte et al. 2012.

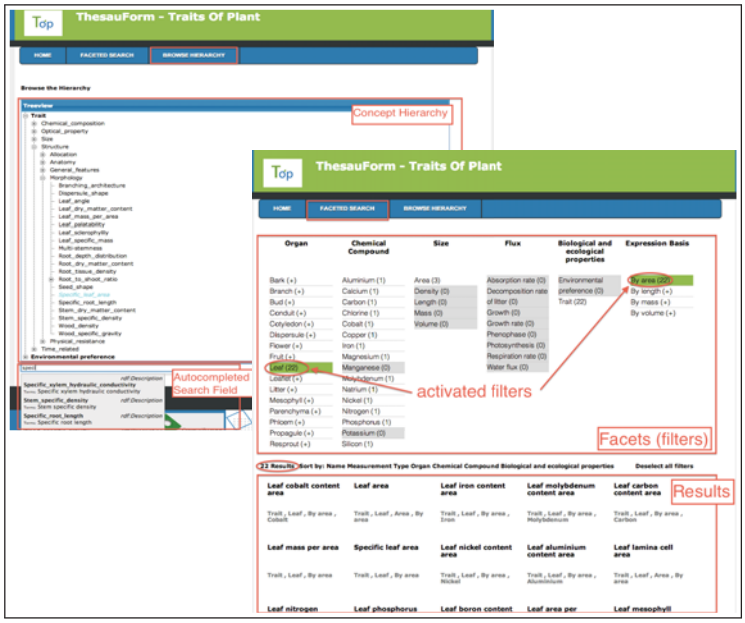

Fig. 1. Semantic search interfaces. On the left, a classic search engine is shown. End-users can use the concept hierarchy tree or an auto-completed field to look for a specific trait. On the right, a faceted search is proposed in order to facilitate and assist the search. The results list is dynamically updated according to the selected filters in the result part.

and over different related domains, specific information for each TOP thesaurus term has been enriched with existing data standard. In Figure 2, the interface displays the results obtained after a query on the TOP thesaurus. The first part of the interface is dedicated to information specific to plant trait, resulting of the collaborative edition of the thesaurus by the community experts. The second part of the interface is about the information from external data sources. For instance, 65157 observations are referenced in the TRY database about Specific Leaf Area.

4.3 Technologies used

We implement a "thin-client/application server" architecture using the J2EE platform, with the system application server being deployed on Apache Tomcat. We used the Jena API to manage the aspects related to the manipulation of the SKOS thesaurus. As we developed a traditional web application, we utilize jquery libraries to support dynamic aspects.

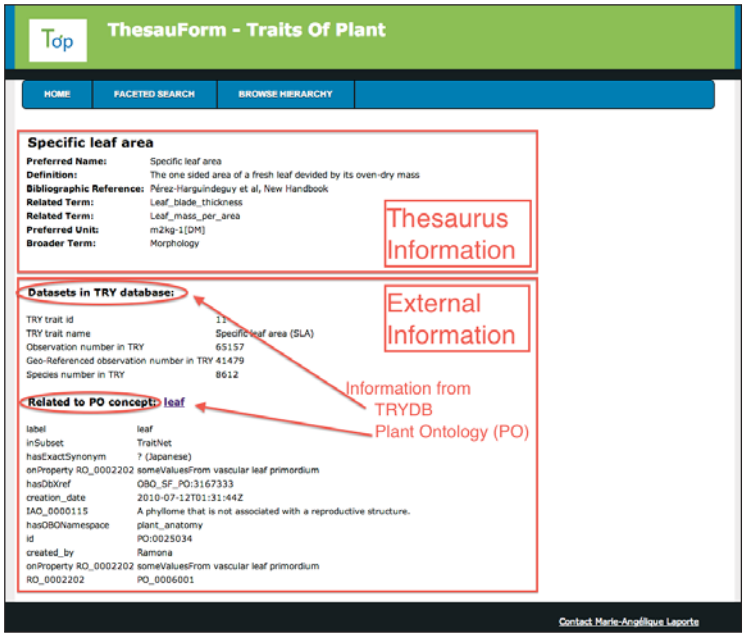

Fig. 2. Trait restitution information interface. Information from both trait thesaurus and external sources is displayed. By now, information from the TRY Database concerning observation number and type, and plant organs or tissues information from Plant Ontology (PO) have been made available.

5 Conclusion and perspectives

Recent studies highlight the crucial need to dispose thesaurus in the field of biodiversity and more precisely in the field of plant diversity [2], [24]. Plant trait research is complex and requires information from different domains to fully exploit plant trait data. Consequently, we propose a complete system designed to the needs of the plant trait community. Such a system provides a tool to build a SKOS thesaurus, assists a community of experts to manage their datasets, and to interconnect them with data and data standards from related communities using the trait thesaurus. We argue that the end-user preferences have to be of prime importance in data access and retrieval. In this context, a faceted search engine demonstrates its full capabilities. Having data sources semantically annotated with TOP thesaurus concepts can benefit from faceted search engine traits and can be used to access disseminated information from heterogeneous data sources. The approach championed in this paper has been to base our work on the continuity of the Open Linked Data initiative .

The impact of the present work is therefore far reaching. First we propose that, just as the molecular biology community has succeeded in during the past twenty years, the functional ecology community has to widely use controlled

vocabularies, thesaurus and ontology, including the TOP thesaurus, in order to describe and annotate their data in the future years. Second, the available data sets have to be made open source. Third, as illustrated by the use case described in this paper and based on mapping approaches with existing controlled vocabularies or ontologies enhancing data interoperability, the data could reveal their huge capabilities. We highlighted numerous relevant ontologies for such a problematic on the NCBO BioPortal. A next step will be to propose more mapping to external resources (both data and controlled vocabularies/ontologies) with the TOP thesaurus. A significant limitation to this kind of approach in an era of Linked Data is to dispose of controlled vocabularies and ontologies compliant with RDF and all the ensuing Semantic Web standards.

References

1. Jones, M.B., Schildhauer, M.P., Reichman, O., Bowers, S.: The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. Annual Review of Ecology, Evolution, and Systematics **37**(1) (December 2006) 519–544
2. Reichman, O.J., Jones, M.B., Schildhauer, M.P.: Challenges and Opportunities of Open Data in Ecology. Science **331**(6018) (February 2011) 703–705
3. Michener, W.K., Jones, M.B.: Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution **27**(2) (February 2012) 85–93
4. Lavorel, S., Garnier, E.: Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. Functional Ecology **16**(5) (October 2002) 545–556
5. Díaz, S., Fargione, J., Chapin, F.S., Tilman, D.: Biodiversity loss threatens human well-being. PLoS biology **4**(8) (August 2006) e277
6. Naeem, S., Bunker, D.: TraitNet: furthering biodiversity research through the curation, discovery, and sharing of species trait data. In: Biodiversity, Ecosystem Functioning, and Human Wellbeing: An Ecological and Economic Perspective. (2009) 281–289
7. Garnier, E., Navas, M.L.: A trait-based approach to comparative functional plant ecology: concepts, methods and applications for agroecology. A review. cef-cfr.ca (Umr 5175) (2011)
8. Violle, C., Navas, M.L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., Garnier, E.: Let the concept of trait be functional! Oikos (116) (2007) 882–892
9. Kattge, J., Ogle, K., Bonisch, G., Diaz, S., Lavorel, S., Madin, J., Nadrowski, K., Nollert, S., Sartor, K., Wirth, C.: A generic structure for plant trait databases. Methods in Ecology & Evolution (2010)
10. Hendry, G., Grime J.P: Methods in comparative plant ecology. Chapman & Hall, London (1993)
11. Cornelissen, J.H.C., Lavorel, S., Garnier, E., Díaz, S., Buchmann, N., Gurvich, D.E., Reich, P.B., Steege, H.T., Morgan, H.D., Heijden, M.G.a.V.D., Pausas, J.G., Poorter, H.: A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. Australian Journal of Botany **51**(4) (2003) 335
12. Knevel, I.C., Bekker, R.M., Bakker, J.P., Kleyer, M.: Life-history traits of the Northwest European flora: The LEDA database. Journal of Vegetation Science **14**(4) (2003) 611–614
13. Berners-Lee, T.: Semantic Web. Conference XML 2000 (2001)

14. T.Berners-Lee: linked data
15. Isaac, A., Summers, E.: SKOS Simple Knowledge Organization System Primer. W3C Technical Report (2008)
16. Walls, R., Cooper, L., Elser, J., Stevenson, D.: The Plant Ontology: A Common Reference Ontology for Plants. wiki.plantontology.org (2010) 2010
17. Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.a., Stevenson, D.W., Smith, B., Preece, J., Athreya, B., Mungall, C.J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T.Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., Jaiswal, P.: The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant & cell physiology **54**(2) (February 2013) e1
18. Laporte, M.A., Mougenot, I., Garnier, E.: ThesauForm—Traits: A web based collaborative tool to develop a thesaurus for plant functional diversity research. Ecological Informatics (May 2012)
19. Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G.F., Hancock, D., Morrison, N., Bruskiewich, R., McLaren, G.: Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. AoB plants **2010** (January 2010) plq008
20. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., Musen, M.a.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research **37**(Web Server issue) (July 2009) W170–3
21. Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., Summers, E.: Key Choices in the Design of Simple Knowledge Organization System (SKOS). eprint arXiv:1302.1224 (2013)
22. Brugman, H., Malaisé, V., Gazendam, L.: A Web Based General Thesaurus Browser to Support Indexing of Television and Radio Programs. 6–9
23. Suominen, O., Viljanen, K., HyvÄnen, E.: User-Centric Faceted Search for Semantic Portals. In Franconi, E., Kifer, M., May, W., eds.: The Semantic Web: Research and Applications. Volume 4519 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2007) 356–370
24. Catapano, T., Hobern, D., Lapp, H., Morris, R.A., Morrison, N., Noy, N., Schildhauer, M., Thau, D.: Recommendations for the Use of Knowledge Organisation Systems by GBIF. Global Biodiversity (2011)

Crop Ontology: Vocabulary For Crop-related Concepts

Luca Matteis[1], Pierre-Yves Chibon[2], Herlin Espinosa[3], Milko Skofic[1], Richard Finkers[2], Richard Bruskiewich[1], Glenn Hyman[3], and Elizabeth Arnaud[1]

[1] Bioversity International
Via dei Tre Denari 472/a
00057 Maccarese (Fiumicino) Rome, Italy
{l.matteis,m.skofic,r.bruskiewich,e.arnaud}@cgiar.org
http://www.bioversityinternational.org/
[2] Wageningen UR Plant Breeding
Wageningen University and Research Centre,
PO Box 386, 6700 AJ, Netherlands
py@chibon.fr, richard.finkers@wur.nl
[3] CIAT, International Center for Tropical Agriculture
Km 17, Recta Cali-Palmira Apartado Areo 6713
Cali, Colombia
{h.r.espinosa,g.hyman}@cgiar.org

Abstract. A recurrent issue for data integration is the lack of a common and structured vocabulary used by different parties to describe their data sets. The Crop Ontology (www.cropontology.org) project aims to provide a central place where the crop community can gather to generate such standardized vocabularies and structure them into ontologies. Having standardized ontologies opens the world of the Semantic Web to data integration between different data providers. Crop Ontology is a community-based project, providing a central place for the creation of crop-related ontologies, but it can also be integrated into third-party tools through its Application Programming Interface, providing retrieval of specific terms or a more generic search functionality for all terms. The ontologies are available in RDF format, described using the OWL and RDFS standards, allowing them to be consumed by popular semantic reasoners. We believe that Crop Ontology will lead to better description of crop-related data, improving collaboration between partners and should serve as an example for other scientific fields.

Keywords: vocabularies, ontologies, Semantic Web, Linked Data, agricultural biodiversity, crops

1 Introduction

Over the last decade there has been a large increase in the number of online vocabularies and ontologies [1]. Search engines such as Google, Yahoo! and Bing, have agreed on a common vocabulary that describes entries in their databases.

This vocabulary is hosted on `http://schema.org`, allowing search engines to be consistent on the meaning of specific concepts. Many other vocabularies exist across the internet, and services such as `http://vocab.cc` allow searching them.

The Linked Data [2] initiative tries to link information across the web using the Semantic Web RDF[4] technology as a basis. This framework enforces the use of URIs[5] for uniquely identifying terms inside a vocabulary or ontology. This initiative has allowed the linkage of data across the web, leading to the construction of a major cloud of information [3].

This cloud however lacks crop-related data. One of the reasons for this, is the lack of standardized vocabularies, which would allow various data provides to describe their data in a consistent manner. Searching for crop terminology on popular ontology search engines[6] websites, shows that very few standards exist in this field.

To build a standardized vocabulary that can be used by different data providers, data providers need to work together. Therefore the Crop Ontology has been built as a community-based project, allowing each member of the community to participate in the building of a vocabulary that matches their needs.

The website was developed as part of a formal Integrated Breeding Platform[7] project of the Generation Challenge Programme[8], to specify global semantic standards for germplasm information management.

2 What is Crop Ontology?

Crop Ontology (`www.cropontology.org`) allows browsing and searching a large database of crop-related terminology, structured per phenotype, breeding, germplasm and trait categories [4–6]. All of this information is freely accessible and downloadable directly through the website. Users can take part in enriching the Crop Ontology database: they can create an account and modify information through a wiki-like system that enables collaboration.

The key feature of this system is that it stores concepts in the form of ontologies. One of the most interesting aspects of building ontologies, instead of simply being a list of descriptors, is that they define relationships between concepts within a specific domain. As useful as this may sound to humans, it becomes even more important for computers. Because it is computers that are capable of understanding what these relationships mean, and can therefore help find information through semantic reasoners [7].

[4] The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. `http://www.w3.org/RDF/`

[5] In computing, a uniform resource identifier (URI) is a string of characters used to identify a name or a web resource. `http://www.w3.org/Addressing/`

[6] Ontology search-engine: `http://lov.okfn.org/dataset/lov/` or `http://swoogle.umbc.edu/`

[7] IBP; `https://www.integratedbreeding.net/`

[8] GCP; `http://www.generationcp.org/`

The ontologies are moderated by semantic experts who help model them, so that they can be consumed by popular semantic reasoners. Moderators of the system make sure everything is done correctly, using good semantic practices. It is important to use standard terminology to build ontologies. OWL [8] and RDFS [9] provide the foundation for these rules, and Crop Ontology uses them extensively.

The simple and easy-to-use interface allows users to browse these concepts through a collapsable tree interface, and search for specific terminology using a powerful free-text search engine. Users can then find concepts and provide feedback when needed. These features allow the direct participation of users in the building process of the ontologies.

3 Features

Crop Ontology aims to create a community of contributors interested in building standard ontologies for crop-related topics. In order to build this community and allow it to perform its goal, a number of features have been implemented: an ontology browser; the possibility to create, extend, and model an ontology; to modify and delete terms; to insert comments; and to programmatically access data through an RDF web service.

3.1 The ontology browser

Browsing is an essential feature of the Crop Ontology website. Users can easily explore the various vocabularies, read descriptions of their terms, and download an RDF version of them. It is simple to find their way through the different types of ontologies, and see the crops available, directly from within the homepage as shown in Figure 1.

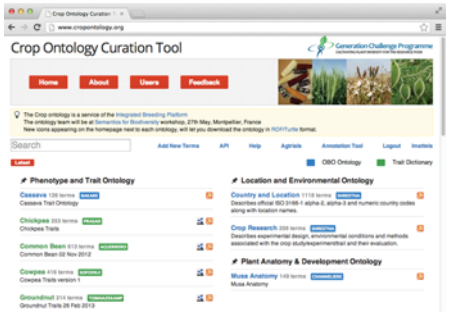

Fig. 1. Homepage of the Crop Ontology website

3.2 Create, extend and model

From the "Create an Ontology" page, as shown in Figure 2, users can immediately start experimenting with a basic interface for building ontologies. Users can create terms directly from within the website, through a dynamic collapsable tree structure. They can insert the name of concepts, and assign basic relationships to each of them, essentially allowing anyone to build a graph through a basic browser-side interface.

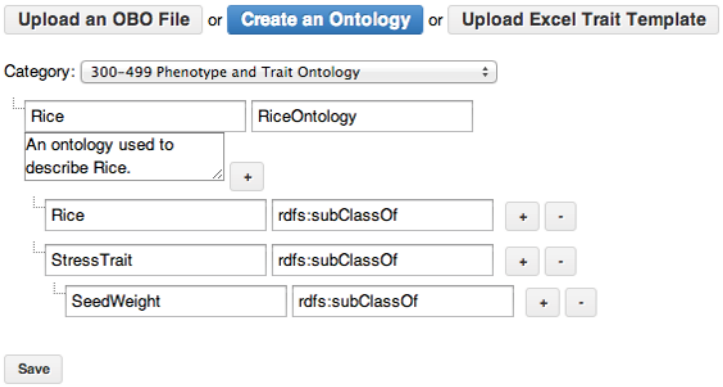

Fig. 2. Web interface for creating ontologies

3.3 Modify and delete

Through the same minimalist interface the system allows also users to modify properties of specific terms. They can insert text in various languages, and upload images that allow them to better describe a concept. Crop Ontology provides simple interface components to allow anybody to modify and extend vocabularies. Figure 3 shows how "action buttons" appear at the right side of each property section, allowing users to quickly identify the action needed to modify or delete a term.

3.4 Leaving comments

Communication is one of the most important parts of community building, so in order for Crop Ontology to build its community of experts, some means of communication between the members is necessary. Under each term, a "comments" section allows users to provide feedback (Fig. 4) directly to the ontology maintainer.

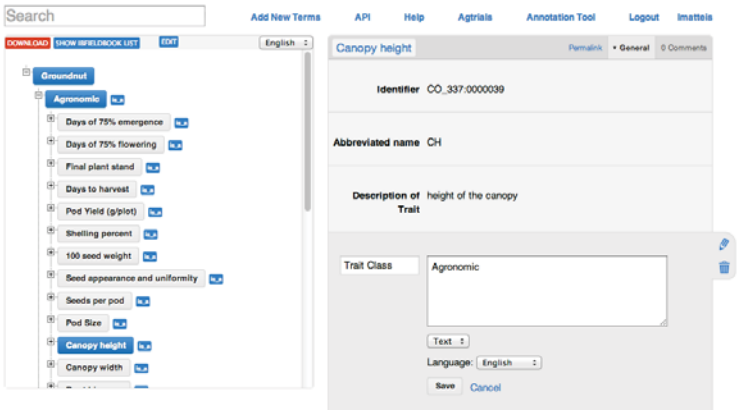

Fig. 3. Edit a term directly from the web-interface

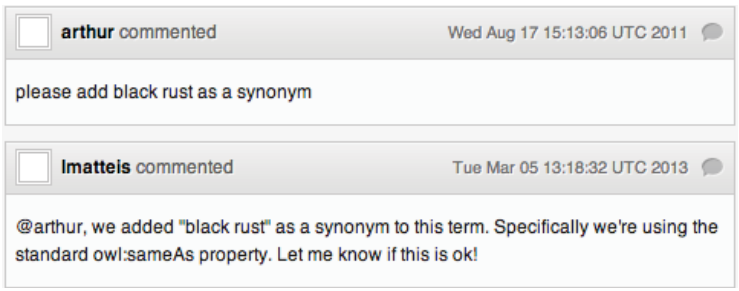

Fig. 4. Comments from the Crop Ontology website

3.5 Graph visualization

Visualizing an ontology as a graph can help the understanding of the relationship between different concepts and how these concepts are structured within their specific domain. An example of an ontology graph is shown in Figure 5.

3.6 RDF support

Crop Ontology decided to adopt the RDF framework. RDF relies on the idea that any piece of information can be described in the form of subject-predicate-object expressions, known as triples. The interesting aspect of the triples is that they are capable of universally storing and linking data: resources are described using URIs, which allows data to be identified and linked in a standard common way, using referenceable resources.

RDFS and OWL are used within the Crop Ontology as they provide standard vocabularies for defining, relating and giving meaning to concepts. By making crop-related data compliant to these standards, they can feed into other data that also use this format, and benefit from them in ways it couldn't otherwise.

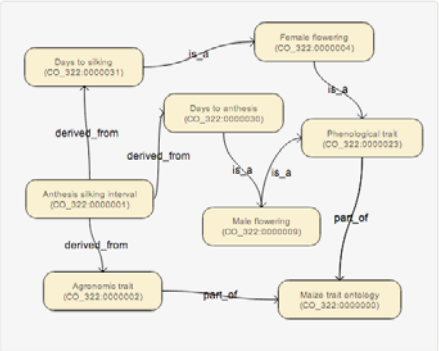

Fig. 5. Graph representation of an ontology

Each URI is structured using the `http://www.cropontology.org/rdf/` namespace, therefore all of the term identifiers are preceded with this URL. Most of the ontologies have initially been modeled using the OBO-Edit[9] software, which generates an OBO file format[10]. Crop Ontology however considers RDF to be a more interoperable format and tries to convert most of the OBO predicates into reasonable RDFS relationships.

RDF also uses the RFC3066 standard for language tags for literals, so this is a built-in feature that the OBO file format standard doesn't support. As most of Crop Ontology's terms are also available in different languages, RDF's multilingual support was very valuable and it allowed for a more natural representation of each concept.

Crop Ontology therefore provides an RDF vocabulary for crop-related data. This means that any system that is managing crop data, can download an RDF format of the ontologies available from the website, and instantly benefit from the work done in defining, linking and giving meaning to these crop-related concepts.

4 Technology

Researchers have greatly benefited from open-source, which creates a collaborative development environment [10]. The Crop Ontology platform therefore was developed from the outset with open-source in mind. By reusing well known libraries and frameworks, the system has been developed on top of a robust underlying structure, which provides greater stability and security. All of the code is publicly available and documented on GitHub[11]:

[9] OBO-Edit is an open source ontology editor written in Java. `http://oboedit.org/`
[10] OBO biological ontology file format. `http://www.geneontology.org/GO.format.shtml`
[11] Online project hosting service. `https://github.com/`

https://github.com/bioversity/Crop-Ontology.
Anybody can use and improve this system, making it a piece of software that others can model to fit their needs.

The ontologies can be downloaded in the popular RDF/Turtle[12] format. This format is well supported by many semantic reasoners such as Apache Jena[13], and it is possible to convert it into other RDF serialization formats if needed.

Google App Engine[14] is also a major component of the Crop Ontology stack. Hosting the application on Google's cloud relieves concerns about the underlying hardware of the computers that are running the software. This gives us more time to concentrate on the development of the product itself, without concerns regarding system administration tasks.

The cloud also provides greater scalability. Many servers are instantiated based on the request load. This essentially makes the system resilient to high-traffic demand, and more resistant against brute-force attacks.

5 Conclusions

Linked Data, and all the technology behind it, is clearly the foundation for data integration of various different information resources. Providing a simple user-interface, such as Crop Ontology, to novice users who are not familiar with all the technologies involved, has proved to be a useful exercise. It has given users the capacity to transform their databases, that were hidden behind personalized schemas, into sharable and linkable resources.

Crop communities are going to continue being involved in the creation of crop-related vocabularies. There are huge numbers of crops that have not been described, and a great deal of information that has not been annotated. The work of bringing more species and more groups into the picture is critical for the continued success of the Crop Ontology. Apart from the Integrated Breeding Platform, many other crop data providers have expressed their interest in us-

[12] Turtle (Terse RDF Triple Language) is a format for expressing data in the Resource Description Framework (RDF) data model. http://www.w3.org/TeamSubmission/turtle/

[13] Jena provides a collection semantic tools and Java libraries. http://jena.apache.org/

[14] Google App Engine is a platform as a service (PaaS) cloud computing platform for developing and hosting web applications in Google-managed data centers. https://developers.google.com/appengine/

ing the Crop Ontology: AgTrials[15], GENESYS[16] and GRIN-Global[17] are in the process of making their data available as RDF resources, with proper linkages to Crop Ontology, allowing it to be linked and discoverable within the Semantic Web.

The system will continue growing with new features also thanks to the open-source community behind it, which constantly feeds the project with fixes and improvements. The future roadmap for the project development includes better integration with richer OWL sublanguages such as OWL DL[18], which allows for greater expressiveness and more complex relationships of the ontologies.

Finally we think that Crop Ontology not only is a useful software system capable of modeling generic ontologies, but in the context of agricultural biodiversity it also provides a meeting ground for various crop communities to discuss and build the next generation of standard crop vocabularies, which are an essential component for the future of biodiversity data management and discoverability.

6 Acknowledgments

The authors would like to thank the data providers who have contributed to submitting data to the Crop Ontology: Peter Kulakow, Bakare Moshood, Sam Ofodile, Ousmane Boukare, Antonio Lopez Montes (IITA); Trushar Shah, Prasad Peteti, Praveen R Reddy, Ibrahima Sissoko, Eva Weltzien, Isabel Vales, Suyah Patil (ICRISAT); Reinhard Simon (CIP); Inge van den Bergh, Stephanie Channeliere (Bioversity International); Mauleon Ramil, Nikki Borgia, Ruaraidh Sackville-Hamilton (IRRI); Alberto Fabio Guerero, Steve Beebe, Roland Chirwa (CIAT). We would also like to thank Rosemary Shrestha and Thomas Hazekamp for providing technical expertise in the field of ontology development, and Arwen Bailey (Bioversity International) for her editorial support. Finally we thank Generation Challenge Programme (GCP) for providing the fund for this collaborative Crop Ontology development and implementation project work.

References

1. Vatant, B., Vandenbussche, P.: `http://lov.okfn.org/dataset/lov/stats/` (2013)

[15] AgTrials is an information portal developed by the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). `http://www.agtrials.org/`

[16] GENESYS is an important and very rich source of information on plant genetic resources diversity of seeds conserved in Genebanks worldwide, crops and crop wild relative material. `http://www.genesys-pgr.org/`

[17] GRIN-Global provides the worlds crop genebanks with a powerful, flexible, easy-to-use global plant genetic resource information management system. `http://www.grin-global.org/`

[18] OWL DL supports those users who want the maximum expressiveness while retaining computational completeness. `http://www.w3.org/TR/2004/REC-owl-features-20040210/#s2.2`

2. Berners-Lee, T.: http://www.w3.org/DesignIssues/LinkedData.html (2006)
3. Heath, T.: http://lod-cloud.net/ (2013)
4. Richard Bruskiewich, Guy Davenport, Tom Hazekamp, Thomas Metz, Manuel Ruiz, Reinhard Simon, Masaru Takeya, Jennifer Lee, Martin Senger, Graham McLaren, and Theo Van Hintum. 2006. The Generation Challenge Programme (GCP)-Standards for Crop Data. OMICS 10(2):215-219
5. Rosemary Shrestha, Elizabeth Arnaud, Ramil Mauleon, Martin Senger,Guy F. Davenport, David Hancock, Norman Morrison, Richard Bruskiewich and Graham McLaren. 2010. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. AoB PLANTS 2010 doi: 10.1093/aobpla/plq008
6. Rosemary Shrestha, Luca Matteis, Milko Skofic, Arllet Portugal, Graham McLaren, Glenn Hyman and Elizabeth Arnaud. 2012. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. Front. Physiol— doi: 10.3389/fphys.2012.00326
7. R. Mishra and S. Kumar. Semantic web reasoners and languages. Artificial Intelligence Review, 2010. DOI 10.1007/s10462-010-9197-3
8. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., PatelSchneider, P.F., Stein, L.A.: OWL Web Ontology Language Reference. Technical Report http://www.w3.org/TR/2004/REC-owl-ref-20040210/, W3C (2004)
9. Brickley, D., R.V. Guha: http://www.w3.org/TR/rdf-schema/ (2004)
10. Gardler, R.: http://news.slashdot.org/story/13/01/29/2252237/how-open-source-could-benefit-academic-research (2013)

A Case-Study of Ontology-Driven Semantic Mediation of Flower-Visiting Data from Heterogeneous Data-Stores in Three South African Natural History Collections

Willem Coetzer[1], Deshendran Moodley[2], and Aurona Gerber[3]

CAIR (Centre for Artificial Intelligence Research), University of KwaZulu-Natal (Durban) / CSIR (Pretoria), South Africa
[1]{w.coetzer@saiab.ac.za} [2]{moodleyd37@ukzn.ac.za}
[3]{agerber@csir.co.za}

Abstract. The domain complexity and structural- and semantic heterogeneity of biodiversity data, as well as idiosyncratic legacy data-creation processes, present significant integration and interoperability challenges. In this paper we describe a case-study of ontology-driven semantic mediation using records of flower-visiting insects from three natural history collections in South Africa. We establish a conceptual domain model for flower-visiting, expressed in an OWL ontology, and use it to semantically enrich the three data-stores. We show how this enrichment allows for the creation of an integrated flower visiting data set. We discuss how this ontology captures both implicit and explicit knowledge, how it can be used to identify and analyze high-level flower-visiting behaviour, and ultimately to construct flower-visiting and pollination networks.

Keywords: biodiversity information, semantic mediation, ontology, plant-insect interactions, pollination

1 Introduction

The challenges of integrating, or making interoperable, distributed, heterogeneous sources of biodiversity- and ecological data have been described [1] [2]. Biodiversity is a complex domain and is no different from other domains in that users encode different definitions of the same concepts [3], which frustrates efforts to integrate data.

We present a case study of three data-stores of flower-visiting insect specimens. All three data-stores consistently contained the names of the plant species, termed *host-plants*, with which both flower-visiting and non-flower-visiting insect specimens were associated. Whereas flower-visiting records were not explicit in most records of two

data-stores, most records of the third data-store contained explicit, easily distinguishable flower-visiting data. To develop a semantic mediation solution, we created the first version of an OWL ontology containing concepts related to flower-visiting and the utilization of flower products, as well as the bearing of pollen by insect vectors. Our work will facilitate the construction of a system to bring about interoperability between distributed and heterogeneous biodiversity data-stores and systems. This will enable biodiversity scientists to more easily extract and analyze the behaviour of flower-visiting insects. Such a system would allow flower-visiting and pollination networks to be automatically assembled and compared.

Outline. In Section 2 we sketch the background against which the need for our study emerged, discuss previous work in biodiversity semantics, and introduce our case-study of interoperability of flower-visiting data. Section 3 begins by describing the domain of flower-visiting and pollination, including our scope, before explaining the process of ontology construction. Expert- and implicit knowledge is highlighted. The usefulness of the concepts in the ontology is discussed in Section 4, by linking data from the data-stores to classes in the ontology. Finally we discuss our approach to a potential solution, including areas where future work is required, and conclude.

2 Background

2.1 Semantics in Biodiversity Informatics

The importance of verifiable specimen-vouchers (i.e. physical preparations such as pinned insects) in museum collections has caused attention to be focused on such specimen information [4]. In recent years *observations* of biodiversity have become important, including observations made by citizen scientists [5]. Both voucher records and observations (collectively termed occurrences) have been subject to the development and adoption of useful standards for publishing and exchanging biodiversity information (the group known as Biodiversity Information Standards (BIS), formerly called the Taxonomic Databases Working Group or TDWG) [6]. One of the BIS standards is the set of terms named the Darwin Core, which contain 'clearly defined semantics that can be understood by people or interpreted by machines, making it possible to determine appropriate uses of the data encoded therein' [7]. The purpose of the Darwin Core terms is to allow biodiversity data to be published and integrated [7].

Biodiversity data are commonly formatted according to the Darwin Core standard and then uploaded to a Global Biodiversity Information Facility (GBIF) participant node (such as the South African Biodiversity Information Facility, SABIF). The data then become discoverable via the GBIF Data Portal, and may be downloaded upon acceptance of conditions. Whereas such database federation has been successful for the sharing of core data attributes (e.g. the Darwin Core categorizes terms as relating to

Occurrence, Event, Location, Identification, Taxon), more specialized data, for example data that record biotic interactions such as parasitism or pollination, are typically omitted because standard terms to describe specific instances of ecological interactions do not yet exist. Currently, shared data therefore fall short of the common phrase 'who did what to whom, where, when, how and why?' because the 'what', 'how' and 'why' are still missing.

The 'Who' and 'To Whom'. The Taxon Concept Schema (TCS) [8] [9], is a standard model to exchange taxonomic information (hence the alternative name 'Taxonomic Concept Transfer Schema'). The TCS is written in XML. More specifically, the TCS allows 'explicit communication of information about Taxon Concepts and their associated names' [8]. A Taxon Concept is a concept or definition of a group, such as a new beetle species, in a taxonomist's mind, which may become published in an article. Several collaborative initiatives aim to define standardized concepts to describe the anatomy and morphology of animals e.g. Hymenoptera [10] or plants [11].

The 'Where' and 'When'. The Darwin-SW Ontology is described as 'an ontology using Darwin Core terms to make it possible to describe biodiversity resources in the Semantic Web' [12]. This is seen as particularly useful for publishing, as Linked Open Data, datasets consisting of Darwin Core terms.

Ecological Semantics. Much work has been done to define concepts used in ecology. Ecological Metadata Language (EML) has a long history of practical application [13] [14], and much work has advanced the use of ontologies [15] [16] to create interoperable systems and to enable the execution of scientific workflows [17] [18].

The need for defining the 'what', 'how' and 'why' of biodiversity information. While the Ecology Ontology and Ecological Networks Ontology [15] contain useful constructs, we found no published, formal definitions of biotic interactions, i.e. concepts that describe specific behaviours representing interactions between individual animals, or between plants and animals. Some preliminary work has been done to extend the Darwin Core standard to broadly include interactions [19] by using terms e.g. `VisitedFlowerOf`, `FlowerVisitedBy`, `NestedIn`, `UsedAsNestBy`. A short list of standard terms was proposed [20] specifically for the interaction, `VisitedFlowerOf`. This list contains the elements: `PollinationEvidence`, `PollenRemoval`, `NectarRemoval`, `OilRemoval` and `FlowerPredation`. Doubt has been expressed as to whether this approach will result in the adequate expression of relationships between specimens or observations.

Semantic mediation in biodiversity informatics. An underlying ontology was used to integrate cereals data from public web databases with data from a local database, allowing molecular characteristics and phenotypic expression to be correlated [37]. While the subject of semantic mediation in biodiversity informatics has been addressed as an architecture component (e.g. [17-18]), few examples of practical applications exist.

2.2 Background to the Case Study

The Quality of Biodiversity Data in South African Museums. South African natural history museums participated in a programme [21] to cleanse and migrate their data to a standard relational database schema and application (Specify Collections Management Software, University of Kansas Biodiversity Institute). Despite having general data of a higher quality, and consistency in schema and syntax, participating researchers of flower-visiting were still unable to easily extract meaningful summaries across data-stores because semantic heterogeneity remained an unresolved challenge. Further work was therefore undertaken with three data-stores that contained data related to collections of flower-visiting insects, namely those of the Albany Museum (AM) in Grahamstown, Iziko Museum (SAM) in Cape Town and Plant Protection Research Institute (SANC) in Pretoria. Table 1 summarizes the data attributes that characterized the data-stores and shows how the word *flower(s)* could be used to distinguish flower-visiting records. The heterogeneity of biodiversity information is evident in Table 1. For example, AM is a specialized flower-visiting data-store because it includes even the colours of visited flowers, and almost all the records are marked with the words 'visit' and 'flower' (also Table 2). On the other hand, SANC contains less-meaningful information for a flower-visiting researcher.

Table 1. Data attributes from the three data-stores. FV = percentage explicit flower-visiting records. Flower-visiting records were distinguished by the *Sampling Method* and *Insect Behaviour* attributes.

	SAM sample data (n=2 094) 3% FV	SANC sample data (n=219) 4% FV	AM sample data (n=21 159) 97% FV
Host Type	host-plant	host-plant	host-plant
Host Taxon	Diascia capensis	Ruschia indecora	Indigofera nigromontana
Sampling Method	**flowers**	swept from **flowering** Acacia albida	hand net
Insect Behaviour	foraging on nectar	[no data]	visiting **flowers**
Flower Colour	[no data]	[no data]	deep pink

3 Ontology Construction in the Domain of Flower-Visiting and Pollination

Various kinds of animals, including arthropods (e.g. insects), birds (e.g. hummingbirds and sunbirds) and mammals (e.g. bats) are well-known *flower-visitors* because they live a life of actively, frequently and consistently seeking out flowers in order to utilize the flowers themselves or their products. The most important flower products are nectar, pollen and oil, which are ingested or collected by the flower-visitors. Insects are important flower-visitors and many insect groups have co-evolved as pollinators of plants.

Pollination is defined with varying granularity. A simple definition reads: 'The transfer of pollen from an anther to a stigma' [22]. Some definitions emphasize that all pollination is ultimately an event (one-step process) because it consists of the act by which pollen is deposited on the pollen-receptive surfaces of a flower (or other reproductive structure such as a cone). In the typical case, pollination (cross-pollination) is a two-step process whereby a vector ('carrier') transfers pollen from the anther of one flower to the stigma of another flower [22]. This is the definition that formed the basis of our domain model, though we did not model the process or event of pollination.

In the study of flower-visiting ecology, pollination may or may not be confirmed in a field setting. Confirmation of pollination requires closely following the flower-visitor and recording its behaviour to see whether it actually transfers pollen onto the stigma. Thus, when ecologists refer to 'pollination' or a 'pollinator', unless otherwise stated, the word is usually used loosely to mean 'inferred pollination' or 'potential pollinator'/'pollen vector' (an organism that carries or transports pollen). Flower-visiting records are the basic currency of pollination ecologists because flower-visiting is easier to observe with high confidence.

Scope. We limited our modelling to angiosperms (flowering plants) that are pollinated by vectors i.e. not by an abiotic medium such as wind or water. We circumscribed as flower-visitors those taxa that belong to the phylum Arthropoda i.e. including the terrestrial groups represented broadly by spiders, millipedes (which mostly inhabit the soil) and insects. Plant galls caused by developing insect larvae, including larvae developing in flower-galls, were excluded from the domain. There was no geographic limitation to our study.

3.1 Concepts used in Domain Modelling: Flower-Visiting and Pollen-Bearing

For the purpose of ontology construction we chose to define the concept of a *flower-visitor* broadly, by interpreting a review of flower-visiting insects [23]. This review clearly included in the concept insects that hid in flowers (e.g. thrips), camouflaged themselves against flowers in order to ambush prey (e.g. mantids) or laid eggs in flowers (e.g. fruit flies). An insect can be a flower-visitor even if it does not ingest or

collect nectar, pollen, oil (with or without terpene fragrance), resin, gum, anthers, ovules, seeds, petals or some other part of the flower or the entire flower.

It is generally accepted that pollen-transfer, both from the anther to a flower-visitor and from the flower-visitor to the stigma is an accidental process.[1] A flower-visitor can become more-or-less covered in pollen, which it may then groom off the surfaces of its body using its tarsi (feet) and mouthparts, and pack into the scopa (hairy patch) on the hind leg, or store on the abdomen or in the crop. The pollen is then taken back to the nest and fed to the young (e.g. social bees) or deposited as nest provision for future young (e.g. solitary bees). Some plants, e.g. orchids and milkweeds, produce a pollinium (plural pollinia), or pollen-mass, borne on a sticky stalk that adheres to the flower-visitor's body. The whole complex including the pollinium and the stalk is called a pollinarium (plural pollinaria).

3.2 Expert- and implicit knowledge

Students of flower-visiting and pollination know implicitly that e.g. an adult beetle or fly or wasp of a certain taxonomic group (e.g. monkey beetles of the tribe Hopliini), or any bee (superfamily Apoidea) has only one reason to be associated with a plant, and that is to visit the plant's flowers, usually to ingest or collect nectar or pollen or other flower products. Many publications list known flower-visiting groups [23].

The importance of implicit knowledge is even more pronounced in the particular case of bees of the genus *Rediviva*, consisting of 26 species that are endemic to South Africa, Lesotho and Swaziland. The females only visit a small number of plant species (about 140 species in 14 genera) whose flowers produce oil to attract these particular bees, or they will visit any number of other plant species whose flowers produce nectar instead of oil [24]. The female bees collect and carry the oil using hairs on their especially-adapted, long front legs, and take the oil back to their nests as provision (i.e. the egg is laid on the oil in the nest and the female that laid the egg then abandons the nest while the larva develops by feeding on the oil). Male *Rediviva* bees only visit flowers that produce nectar, which, like the females that visit 'nectar plants', they ingest to sustain themselves. A 'nectar-plant' could be any flowering plant species, in the area that the bee frequents, that happens to have nectar in its flowers at the time. Among all the specimen records in the SANC data-store that were created during the course of preparing two seminal articles on the famous *Rediviva* oil-collecting bees of southern Africa, the words 'visit', 'flower' or 'oil' do not occur once. The reason for this was probably related to the need for critical information to fit onto a small specimen label. No information was lost within the museum because an expert only needs to know the sex of the adult bee specimen and the plant species name to know whether a *Rediviva* bee was collecting nectar or oil, and that it was visiting flowers[25] [26]).

[1] Fig-wasps seem to undertake an intentional pollination ritual [36].

3.3 The Flower-Visiting and Pollen-Bearer Ontology

In this section we describe the semantic analysis and ontology construction process we followed to create the OWL ontology using Protégé [27]. Both bottom-up (i.e. from the data) and top-down ontology construction approaches (i.e. from literature and discussions with experts) were employed. We re-used concepts from the Plant Ontology [11] where possible. In modelling flower-visiting we made extensive use of the `Role` concept as defined in BFO (the Basic Formal Ontology) [28]. Examples of roles include the role of a person as a surgeon or the role of a chemical compound in an experiment. We created –Role concepts for the activities associated with flower visitors , and created an Object Property, `participates_in` (inverse: `participated_in_by`); thus a `FlowerVisitor` `participates_in` `some` `FlowerVisitorRole`. The `–Role` taxonomy is depicted in Figure 1.

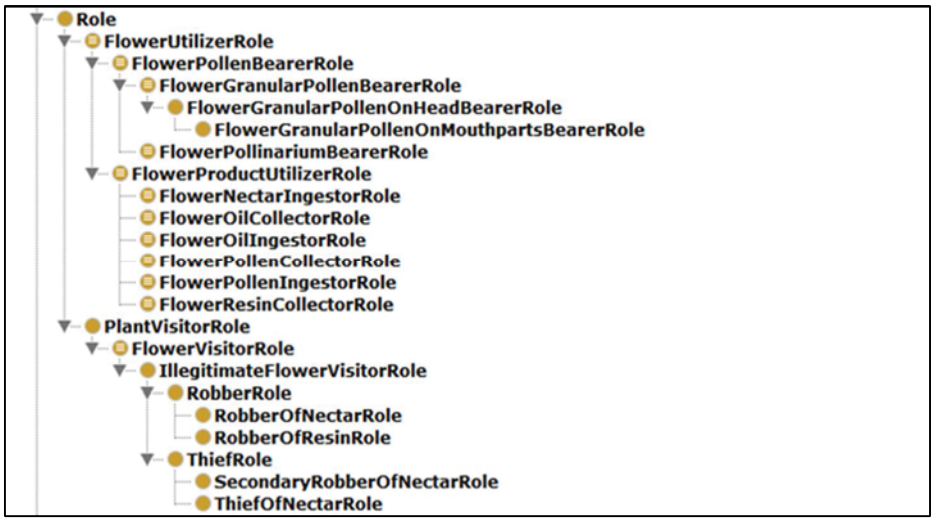

Fig 1. The roles (concepts) in the asserted class hierarchy as displayed in Protégé 4.2

3.4 The `FlowerVisitorRole`

Our objective was to make interoperable heterogeneous records of *flower-visitors*, which are generally organisms that utilize flowers. We therefore created the object property, `utilizes` (inverse: `utilized_by`), and defined the necessary condition for the class `FlowerVisitorRole`:

`utilizes some WholePlant`

This means that an organism on a severed flower lying on the ground, or in a flower arrangement, cannot be a `FlowerVisitor`.

The necessary and sufficient conditions for the class, `FlowerVisitorRole`, are either:

```
A: (utilizes some FlowerMechanicalSupport)
  or (utilizes some FlowerSpace)
  or (utilizes some FlowerTissue)
  or (utilizes some FlowerProduct)

or

B: (participates_in some PlantVisitorRole)
   and (member_of some FlowerVisitingGroup)

or

C: (bears some Pollen) or (bears some Pollinarium)
```

In Section A, `utilizes` some `FlowerMechanicalSupport` could mean alighting on a flower, `utilizes` some FlowerSpace could mean inserting the proboscis into the flower or hiding in the flower. `utilizes` some `FlowerTissue` could mean laying an egg inside the tissue or eating the tissue. `utilizes` some `FlowerProduct` could mean ingesting or collecting nectar or pollen. This class will therefore include individuals that are incidental flower-visitors (e.g. spiders) as well as highly specialized pollen-collectors (e.g. bees).

Section B in the above class definition states that a condition for an organism that `participates_in` the `FlowerVisitorRole` is that it `utilizes some WholePlant and is a (member_of some FlowerVisitingGroup)`.

We created the object property, `bears` (inverse: `borne_by`), meaning to 'have on (the outside of the body)', as in 'the bee's abdomen bears pollen'. This object property was used, in Section C above, to assert that a condition for an organism that `participates_in` the `FlowerVisitorRole` is that it `bears Pollen` or `bears` at least one `Pollinarium`.

3.5 The `FlowerUtilizerRole` and descendent classes, including implicit knowledge of *Rediviva* bees

It was asserted that a condition for the `FlowerUtilizerRole` is `((utilizes some FlowerMechanicalSupport) or (utilizes some FlowerSpace)or (utilizes some FlowerTissue) or (utilizes some FlowerProduct))`. This means that `FlowerUtilizerRole` is equivalent to `FlowerVisitorRole`.

We specialized the object property, `utilizes`, into the object properties, `ingests` (inverse: `ingested_by`) and `collects` (inverse: `collected_by`).

We defined a `FlowerProduct` to be the class subsuming the class (`FlowerSecretion` or `Pollen` or `Pollinarium`). The class `FlowerSecretion` subsumed the class (`FlowerGum` or `FlowerNectar` or `FlowerOil` or `FlowerResin`).

The `FlowerUtilizerRole` was specialized into `FlowerProductUtilizerRole and FlowerPollenBearerRole`. More specifically, if an individual `utilizes` (`ingests or collects`) some `FlowerProduct`, that is sufficient to mean that it `participates_in` the `FlowerProductUtilizerRole`.

An individual that `(bears some Pollen) or (bears some Pollinarium)` sufficiently meets the condition for the `FlowerPollenBearerRole`. If an organism actively ingests or collects pollen, some pollen will invariably remain on its body after grooming and packing into the scopa. A necessary condition of the `FlowerPollenIngestorRole` and the `FlowerPollenCollectorRole` is therefore: `bears some Pollen`. Figure 2 depicts two parts of the inferred class hierarchy: `FlowerProductUtilizer` and sub-classes, as well as detail of the `FlowerPollenCollector` class hierarchy. The classes in Figure 2 are sub-classes of `Organism`. These classes `participate_in` the `–Role` classes depicted in the taxonomy in Figure 1.

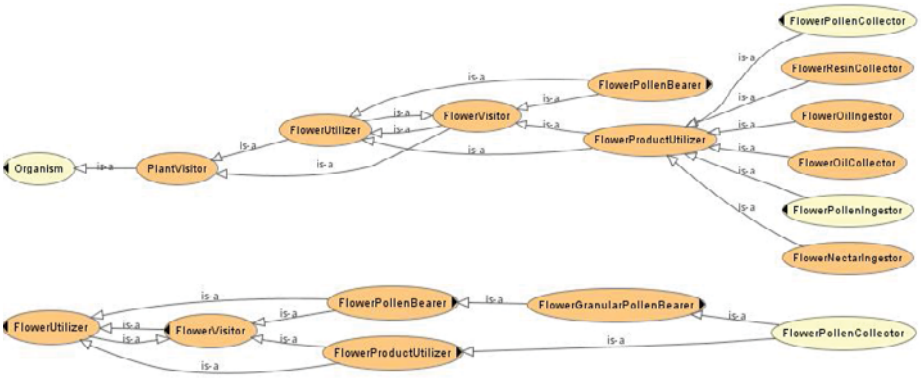

Fig. 2. It is asserted that a `FlowerPollenBearer` need not be a `FlowerProductUtilizer`, but an organism may be both a `FlowerPollenBearer` and a `FlowerProductUtilizer` because these classes are not disjoint. This successfully models active pollen-collecting and pollen-ingesting, which necessarily result in passively bearing pollen.

The conditions that are sufficient for membership in the `FlowerOilCollector` class are as follows: `((participates_in some FlowerOilCollectorRole)) or ((participates_in some OilPlantVisitorRole) and`

`(member_of some FlowerVisitingGroup) and (has_sex only Female) and (part_of some RedivivaGenus))`.

This means that a `FlowerOilCollector` can either be observed directly (`collects some FlowerOil`) or its presence can be inferred (e.g. in the SAM data-store) from the facts that an 'oil plant' (with flowers that secrete oil, not nectar) was visited, the insect was a female and it was a species in the genus *Rediviva*.

3.6 The `IllegitimateFlowerVisitorRole` and sub-classes

With reference to Figure 1, the concept of 'illegitimately' visiting flowers (i.e. by definitely avoiding coming into contact with the anthers, and therefore never becoming a `FlowerPollenBearer`) is frequently encountered in the flower-visiting literature, and we therefore included this in our ontology. Robbers, which damage the petals (e.g. by biting a hole in the petal to access the nectar), are distinguished from thieves, which inflict no petal damage. A secondary robber obtains nectar through the hole made by a primary robber [29].

4 Linking the Ontology to Existing and Future Data

The class, `FlowerUtilizer` (Section A of the definition of the `FlowerVisitorRole`) therefore represents records resulting from the observations of a generalist scientist who may record an organism generally utilizing a flower by e.g. sitting on, or flying around and feeding from (visiting), a flower. In the AM data-store a small number of records were classified as members of the class `FlowerUtilizer` (Table 2).

Table 2. Examples of the class FlowerProductUtilizer in the AM data-store

# records	Behaviour	Class
137	Visiting extrafloral nectaries	PlantVisitor
95	On foliage	PlantVisitor
8	On stem of plant	PlantVisitor
20135	Visiting flowers	FlowerProductUtilizer
380	In flowers	FlowerUtilizer
22	On flowers	FlowerUtilizer
16	Sheltering in flower	FlowerUtilizer
8	In copula on flowers	FlowerUtilizer

The vast majority of records, however, were instances of the class, `FlowerProductUtilizer`. An expert in the study of flower-visitors would record a flower-visitor to be an instance of the class `FlowerProductUtilizer` (i.e. specifically ingesting or collecting nectar or pollen). Importantly, this observation can be made by an expert observing an insect that has not even touched a flower. The expert is able

to classify the organism into a specific taxonomic group, and to remember how previous individuals in this specific group have behaved (i.e. they *visited* flowers, which is a shorter way of recording that they ingested or collected nectar or pollen), and to know that newly observed individuals of the same group are unlikely to behave differently. The predominance of records of the `FlowerProductUtilizer` class therefore reflects the predominance of bees and pollen wasps in this data-store, which is due, in turn, to the development of the careers of the specialists who built the specimen collection. It is therefore not surprising that the biodiversity information in the AM data-store is richer than the information in the other data-stores.

4.2 Data in the SAM and SANC data-stores

Ninety-seven per cent of the records in the SAM data-store, and 96% of the records in the SANC data-store, were instances of the class `FlowerVisitor`, a term that is less meaningful than `FlowerUtilizer` or `FlowerProductUtilizer`. A small number of records in the SAM data-store were instances of sub-classes of the class `FlowerProductUtilizer`. Some of these are shown in Table 3.

Table 3. Examples of the class FlowerProductUtilizer in the SAM data-store

# records	Behaviour	Class
1	Collecting pollen on yellow flowers.	FlowerPollenCollector
1	Patrolling Corymbium. With pollenaria.	FlowerPollinariumBearer
1	Feeding on Brunia laevis pollen	FlowerPollenIngestor
1	Foraging on nectar of Euphorbia flowers.	FlowerNectarIngestor
1	Taking resin from Dalechampia capensis.	FlowerResinCollector

Section C of the definition of the `FlowerVisitorRole` (i.e. a `FlowerPollenBearer`) is of particular, current interest. If an organism is seen to bear pollen or a pollinarium, DNA barcoding can be used to identify [30] the plant species that produced the pollen. This is a very important step in the study of flower-visiting because it means that it will no longer be necessary to observe a `FlowerPollenBearer`, either in any physical association with a plant or flower, or actually ingesting or collecting pollen, to know:

1) That it must be a `FlowerUtilizer` (but not necessarily a `FlowerProductUtilizer`) and therefore a `FlowerVisitor`;
2) The list of plant species which it has recently visited, utilized and borne pollen from.

5 Discussion and Conclusion

We have shown how implicit domain knowledge about flower visitors can be represented in an ontology for use in semantic enrichment of, and semantic mediation between, heterogeneous data sources.

Researchers of flower-visiting need to summarize data into lists of insect species and the plant species whose flowers those insects visit, and which they probably pollinate. These lists usually form the basis of further work involving the modelling of flower-visiting networks (which are useful in community ecology), and, more specifically, pollination networks (e.g. [31]). In an applied study the ultimate objective may be to compare the characteristics [32] of pollination networks across space or through time e.g. to estimate the effect, on pollination, of habitat transformation [33] or global change.

Clearly, systems used to capture and manage specimen data are not designed to capture the background knowledge required to access the rich, and often implicit, information associated with these records. This knowledge is usually held by the curator or scientists who generated the records. This becomes more pronounced for biodiversity researchers accessing a network of locally controlled and heterogeneous biodiversity databases. A significant barrier to data integration and analysis will therefore be removed if knowledge can be explicitly represented within the system. For example, illegitimate flower-visitor species must be excluded from the process of assembling a pollination network.

In our current ontology we assumed that there are no exceptions of a `KnownFlowerVisitingGroup`. This is an area where future work is needed because the semantic representation of exceptions, or defeasibility with current OWL ontologies, is problematic. One of these exceptions is a particular Afrotropical bee species which is an obligate raider of other bees' nests and therefore has no need to, and never does, visit flowers. Yet bees are the most important group of flower-visiting insects. Such exceptions will need to be carefully modelled to prevent the possibility of drawing incorrect inferences.

While the ontology described above can certainly facilitate the creation of a semantically rich flower-visiting data set, it still falls short of capturing uncertain and vague biotic interactions associated with flower-visiting occurrences. Probabilistic graphs such as Bayesian Networks are better able to deal with uncertain causal relations, especially when there is uncertainty and vagueness [34]. The combination of ontologies and Bayesian networks has recently been explored in the earth observation domain within the Sensor Web Agent Platform (SWAP) [35]. In SWAP sensor observations from heterogeneous sensor data-stores are semantically enriched with OWL ontologies and used to populate Bayesian networks to determine the probability of the occurrence of abstract physical earth observation phenomena.

The next step in our semantic mediation system will be to adapt the SWAP [35] approach and construct a Bayesian network that describes the causal relations between plant-visiting events, flower-visiting events, pollen transfer events and pollination events. These events will be defined using concepts from the flower-visiting ontology. In this way semantically enriched observations from the three data-stores can be used as proxies to determine the probabilities of the occurrence of flower-visiting and pollination events.

Acknowledgement

With gratitude we acknowledge the JRS Biodiversity Foundation (http://www.jrsbdf.org/) for financial support of the research presented in this paper through a 2011 grant for *Improvement and Integration of Pollinator Biodiversity Information in Africa.*

References

1. Johnson, N.F.: Biodiversity Informatics. Annual Review of Entomology. 52, 421–38 (2007).
2. Jones, M.B., Schildhauer, M.P., Reichman, O.J., Bowers, S.: The New Bioinformatics: Integrating Ecological Data From the Gene to the Biosphere. Annual Review of Ecology Evolution and Systematics. 37, 519–544 (2006).
3. Deans, A.R., Yoder, M.J., Balhoff, J.P.: Time to Change How We Describe Biodiversity. Trends in Ecology & Evolution. 27, 78–84 (2011).
4. Bisby, F.A.: The Quiet Revolution: Biodiversity Informatics and the Internet. Science. 289, 2309–2312 (2000).
5. Silvertown, J.: A New Dawn For Citizen Science. Trends in Ecology & Evolution. 24, 467–471 (2009).
6. Biodiversity Information Standards, http://www.tdwg.org/.
7. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D.: Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE. 7, e29715 (2012).
8. Kennedy, J., Hyam, R., Kukla, R., Paterson, T.: A Standard Data Model Representation for Taxonomic Information. Omics, a Journal of Integrative Biology. 10, 220–230 (2006).
9. Hyam, R., Kennedy, J.: Taxon Concept Schema – User Guide. Unpublished Report. 28 pp. (2005).
10. Yoder, M.J., Mikó, I., Seltmann, K.C., Bertone, M.A., Deans, A.R.: A Gross Anatomy Ontology For Hymenoptera. PloS one. 5, e15991 (2010).
11. The Plant Ontology Consortium: The Plant Ontology™ Consortium and Plant Ontologies. Comparative and Functional Genomics. 3, 137–142 (2002).
12. Webb, C., Baskauf, S.: Darwin-SW: Darwin Core Data for the Semantic Web.
13. Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G.: Nongeospatial Metadata for the Ecological Sciences. Ecological Applications. 7, 330–342 (1997).

14. Johnson, J.C., Christian, R.R., Brunt, J.W., Hickman, C.R., Waide, R.B.: Evolution of Collaboration within the US Long Term Ecological Research Network. BioScience. 60, 931–940 (2010).
15. Williams, J.R., Martinez, N.D., Golbeck, J.: Ontologies for Ecoinformatics. Web Semantics: Science, Services and Agents on the World Wide Web. 4, 237–276 (2006).
16. Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F.: An Ontology for Describing and Synthesizing Ecological Observation Data. Ecological Informatics. 2, 279–296 (2007).
17. Michener, W.K., Beach, J.H., Jones, M.B., Ludäscher, B., Pennington, D.D., Pereira, R.S., Rajasekar, A., Schildhauer, M.: A Knowledge Environment for the Biodiversity and Ecological Sciences. Journal of Intelligent Information Systems. 29, 111–126 (2007).
18. Michener, W.K., Jones, M.B.: Ecoinformatics: Supporting Ecology as a Data-Intensive Science. Trends in Ecology & Evolution. 27, 85–93 (2012).
19. De Giovanni, R., Cartolano, E., Giannini, T., Saraiva, A., Pizzigatti, P.: Darwin Core Interaction Extension Concept List, http://wiki.tdwg.org/twiki/bin/view/DarwinCore/InteractionExtension.
20. De Giovanni, R., Cartolano, E., Giannini, T., Saraiva, A., Pizzigatti, P.: Darwin Core Interaction Extension: Pollination Extension Concept List, http://wiki.tdwg.org/twiki/bin/view/DarwinCore/PollinationExtension.
21. Coetzer, W., Gon, O., Hamer, M., Parker-Allie, F.: A New Era for Specimen Databases and Biodiversity Information Management in South Africa. Biodiversity Informatics. 8, 1–11 (2012).
22. Raven, P.H., Evert, R.F., Eichhorn, S.E.: Biology of Plants. Worth Publishers, Inc., New York (1986).
23. Kevan, P.G., Baker, H.G.: Insects as Flower Visitors and Pollinators. Annual Review of Entomology. 28, 407–453 (1983).
24. Pauw, A.: Floral Syndromes Accurately Predict Pollination by a Specialized Oil-Collecting Bee (Rediviva peringueyi, Melittidae) in a Guild of South African Orchids (Coryciinae). American Journal of Botany. 93, 917–926 ST – Floral syndromes accurately predict (2006).
25. B Whitehead, V., E Steiner, K.: Oil-collecting Bees of the Winter Rainfall Area of South Africa. Annals of The South African Museum. 108, 143–277 (2000).
26. Whitehead, V.B., Steiner, K.E., Eardley, C.D.: Oil Collecting Bees Mostly of the Summer Rainfall area of Southern Africa (Hymenoptera: Melittidae: Rediviva). Journal of the Kansas Entomological Society. 81, 122–141 (2008).
27. Horridge, M.: A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools Edition 1.3, (2011).
28. Arp, R., Smith, B.: Function, Role, and Disposition in Basic Formal Ontology. Nature. 2, 1–4 (2008).
29. Murphy, C.M., Breed, M.D.: Nectar and Resin Robbing in Stingless Bees. American Entomologist. Spring, 36–44 (2008).
30. Hebert, P.D.N., Cywinska, A., Ball, S.L., DeWaard, J.R.: Biological identifications through DNA Barcodes. Proceedings of the Royal Society B: Biological Sciences. 270, 313–321 (2003).

31. Dupont, Y.L., Padrón, B., Olesen, J.M., Petanidou, T.: Spatio-Temporal Variation in the Structure of Pollination Networks. Oikos. 118, 1261–1269 (2009).
32. Kaiser-Bunbury, C.N., Muff, S., Memmott, J., Müller, C.B., Caflisch, A.: The Robustness of Pollination Networks to the Loss of Species and Interactions: A Quantitative Approach Incorporating Pollinator Behaviour. Ecology Letters. 13, 442–452 (2010).
33. Valdovinos, F.S., Ramos-Jiliberto, R., Flores, J.D., Espinoza, C., López, G.: Structure and Dynamics of Pollination Networks: The Role of Alien Plants. Oikos. 118, 1190–1200 (2009).
34. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, Englewood Cliffs, NJ (2003).
35. Moodley, D., Simonis, I., Tapamo, J.: An Architecture for Managing Knowledge and System Dynamism in the Worldwide Sensor Web. International Journal of Semantic Web and Information Systems: Special issue on Semantics-enhanced Sensor Networks. Internet of Things and Smart Devices. 8, 64–88 (2012).
36. Wiebes, J.T.: Co-evolution of Figs and Their Insect Pollinators. Annual Review of Ecology and Systematics. 10, 1–12 (1979).
37. Sala, A., Bergamaschi, S.: A Mediator Based Approach to Ontology Generation and Querying of Molecular and Phenotypic Cereals Data. International Journal of Metadata, Semantics and Ontologies. 4(1/2), 85-92 (2009).

Flexible Scientific Data Management for Plant Phenomics Research

Peter Ansell[1], Robert Furbank[2], Kutila Gunasekera[1], Jianming Guo[2], David Benn[3], Gareth Williams[3], Xavier Sirault[2]

[1] eResearch Group, School of Information Technology and Electronic Engineering, University of Queensland, Brisbane, Australia
[2] CSIRO Plant industry, High Resolution Plant Phenomics Centre, Canberra, Australia
[3] CSIRO IM&T Advanced Scientific Computing and Research Data Services, Melbourne, Australia

Abstract. In this paper, we expand on the design and implementation of the Phenomics Ontology Driven Data repository [1] (PODD) with respect to the capture, storage and retrieval of data and metadata generated at the High Resolution Plant Phenomics Centre (Canberra, Australia). PODD is a schema-driven Semantic Web database which uses the Resource Description Framework (RDF) model to store semi-structured information. RDF allows PODD to process information about a range of phenomics experiments without needing to define a universal schema for all of the different structures. To illustrate the process, exemplar datasets were generated using a medium throughput, high resolution, three-dimensional digitisation system purposely built for studying plant structure and function simultaneously under specific environmental conditions. The High Performance Compute (HPC), storage and data collection publication aspects of the workflow and their realisation in CSIRO infrastructure are also discussed along with their relationship to PODD.

Keywords: eResearch, Semantic Web, RDF, OWL, Data collection citation, BagIt, Data Access Portal

1 Introduction

Since the genomics era, biology has become a data-driven science. Advances in robotics, automation and imaging, in combination with high performance computing have permitted the rapid production of large and complex biological datasets. Currently, high volumes of heterogeneous image data, physiological and morphological measurements are being acquired by a range of new phenotyping platforms located in purpose built phenomics centres across the world. These large datasets of phenotypic characteristics such as growth rate, plant architecture, photosynthetic performance, yield must be stored and correlated with genotypes. These factors provide evidence of genetic variation in natural and derived genetic populations (e.g. germplasm collections, association genetic

panels, recombinant inbred lines). They also enable a deeper understanding of the dynamic relationship between phenotype, genotype and environment which is necessary to continue delivering the increase in productivity necessary for feeding the world.

The vast array of phenotypic data collected from a variety of phenomics platforms must be combined with metadata explaining how the raw data was collected. This combination of raw data and metadata are then delivered to a range of analysis pipelines, which transform the raw data into aggregated multi-phase datasets, each phase representing a new aggregation or inference from the original raw data. This reduction process converts the raw multi-dimensional data into information which is conceptually interpretable by a human being, i.e. new knowledge. The additional metadata describing the steps taken are recorded to give context to the data.

To make sense of this large amount of information, sophisticated storage, archiving, searching and analysis capabilities are required. To date solutions to this problem have been handled essentially by private companies, and no suitable solution exists in the public domain. Lack of systems, both to manage linked metadata, and controlled vocabularies to describe plant growth and experimental conditions, have severely hampered sharing of plant phenomics data, comparison of results between laboratories and the capacity to carry out meta-analysis of existing data sets.

Thus, to support publicly-funded phenomics activities in Australia, the Phenomics Ontology Driven Data repository (PODD) has been developed as a repository for data produced by the variety of plant imaging and phenotyping platforms available at the High Resolution Plant Phenomics Centre, as well as for recording the contextual metadata associated with plant genotypes, treatments and environmental conditions [1].

In this paper, we describe the workflow management that the High Resolution Plant Phenomics Centre (HRPPC) has implemented for keeping track of its phenomics data, metadata and experimental processes. This complex challenge was addressed by building a multi-disciplinary group of information technology experts and embedding users of phenomics technologies into it. The result of the approach is a state of the art computational and data mining environment, optimised for data access, data discovery and data sharing, which also provides the flexibility for linking genomic information through the use of RDF triples. In this context, we also describe the role of the CSIRO Data Access Portal (DAP) [2] to annotate and store raw and processed datasets. DAP also provides long term secure storage for data collections and the ability to search for, control access to, and cite them via Digital Object Identifiers. PODD manages the mapping of collections located in DAP to PODD projects, providing for the storage of large images and documents unsuited to RDF databases. Figure 1 shows the relationship between components and key data flows.

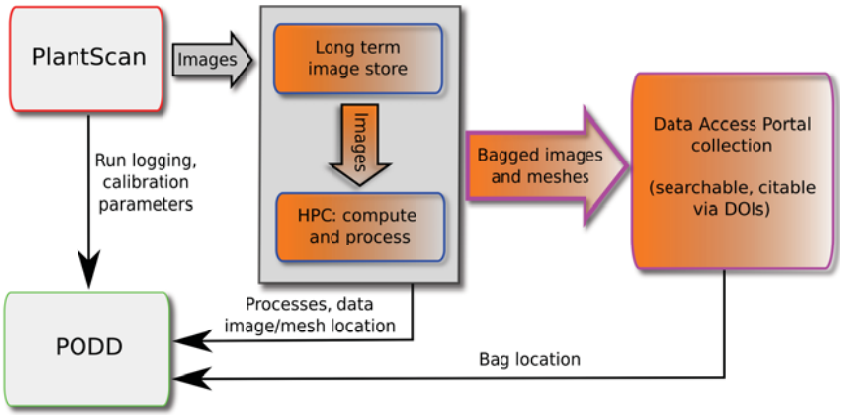

Fig. 1. HRPPC component relationships and data flow

2 Phenomics Ontology Driven Data repository

2.1 Semantic science for phenomics data management

Scientists have focused on including semantics into datasets, typically using the foundations of RDF and OWL, from two main directions. Some focus on defining ontologies based on hierarchies of scientific concepts and properties, while others have focused on mapping complex scientific datasets to RDF using syntax transformations without initially defining the semantic meaning of the results. In reality, most efforts fall somewhere in the middle, with ontological annotations attached to some data points while other nearby data points are syntactically represented using RDF, without links to ontologies of scientific concepts.

Increasingly however, providers of scientific datasets are focusing on enhancing their datasets using curated scientific concepts from ontologies. For example, scientists have used the Gene Ontology [3] to link well known concepts to represent common elements across genomics datasets, while the Plant Ontology [4] allows the description of plant based datasets.

2.2 Redesign of the Phenomics Ontology Driven Data repository

The PODD repository relies on semantic web technologies to manage phenomics data and metadata. Although both ontologies and mappings are essential, in PODD it was necessary to build the system with a relaxed ontological vocabulary. This enables scientists to sparsely populate their datasets and sparsely link to community defined upper ontologies as necessary. This allows scientists to continue to maintain projects containing curated scientific concepts alongside raw experimental data. The PODD repository was redesigned based on an evaluation of the original software [1] that found it was not able to scale sufficiently

to suit the HRPPC needs due to design and implementation deficiencies. The major design differences to the software implemented by [1] are that projects are no longer the only supported top object type, and projects are not stored in multiple parts, as that approach was not able to scale as was originally hypothesised.

A PODD project in PlantScan™ contains top level branches describing the various parts of a scientific project. These include a branch for raw data, along with separate branches for results, analysis, and publications related to the project. In the case of raw data, the semantics are not necessarily clear and are not easily defined by the automated platforms collecting the data. The scientist may later semantically link the data with results, conclusions, and external ontologies. For example, a scientist may annotate the data objects representing images of a plant with a link to a trait that is defined in the Plant Ontology. They may also annotate the image with a link to a trait that is defined inside of the project, such as when the trait is novel and not represented in a community ontology.

2.3 Semantic validation

PODD validates scientific project descriptions using independently configurable constraints based on OWL (Web Ontology Language) ontologies. Although PODD currently solely supports OWL for constraint verification, it could be easily extended in other cases to use different systems such as N3, RDFS, SPARQL, or SPIN as rules languages [5].

OWL is used to determine whether projects are both internally consistent, with all objects having an explicit RDF type, and whether they are consistent with the ontologies that they import. For example, any OWL object property that has been defined to link from image acquisition runs to images defines the provenance of an image.

General scientific properties and phenotype specific properties are defined in optional extension ontologies as illustrated in Figure 2. These are used by scientists to annotate their projects with concepts specific to their field, without requiring other scientists using the same PODD installation to use phenotype properties to annotate their projects.

3 CSIRO Data Access Portal

CSIRO's Research Data Service (RDS) has developed the Data Access Portal (DAP), an open source web application that enables research data to be discovered, managed and shared. [2]

Researchers can describe a data collection, deposit data, choose a license, and add attribution details. Access to a collection's description and/or data can be restricted to CSIRO or a set of individuals (within CSIRO or partner organisations) or it can be made public, becoming searchable by anyone via the Internet. In the case where a collection and its data are public, a Digital Object

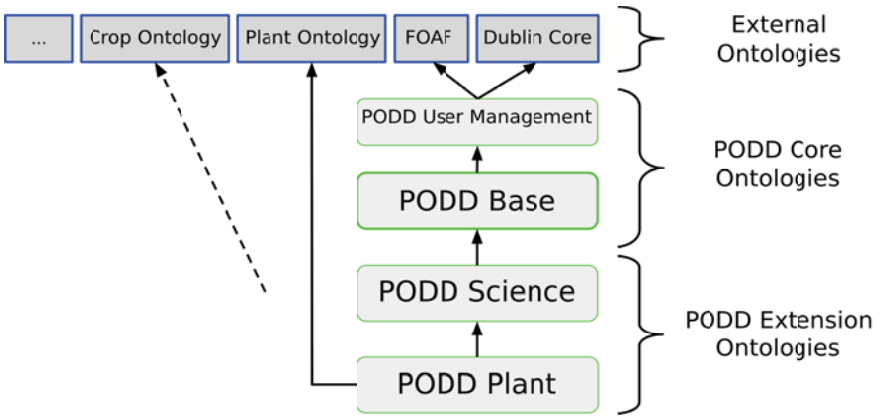

Fig. 2. PODD ontology hierarchy

Identifier (DOI) is issued and can be used to formally cite the collection in a publication.

4 The PlantScan™ digitisation platform

4.1 BagIt

BagIt is defined by an Internet Engineering Task Force (IETF) document as an "hierarchical file packaging format for storage and transfer of arbitrary digital content"[6]. A payload manifest details content and MD5 or SHA hashes for content integrity verification. Data file related metadata can be stored in pre-defined files as key-value pairs.

For PlantScan™, file-level metadata includes plant barcodes, batch numbers, and plant type, although the BagIt specification does not mandate a particular archiving strategy, with the focus being upon the directory structure, special files, and integrity checking. BagIt-conforming tools [7] [8] were assessed and where necessary, improvements were implemented and tested to ensure that the tools were fit for purpose in the CSIRO Advanced Scientific Computing (ASC) HPC environment.

4.2 Bag preparation for a DAP collection

CSIRO ASC shared facilities [9] are used to process the raw PlantScan™ data to derive data products (meshes). Raw data and meshes are collected using the BagIt format [6] and stored in the ASC archival system. ASC High Performance Compute (HPC) hosts (systems with high processor count and large memory) are taken advantage of to create and verify bags more rapidly than would be possible on conventional computer systems. CSIRO's HRPPC makes use of DAP

to store collections of PlantScan™ raw images and processed mesh data as bags. Currently, one bag is equivalent to a single batch scanned on the PlantScan™ local software system, which usually means the same kind of plant with different genotypes scanned under one experiment configuration profile.

Raw data from PlantScan™ local storage (HRPPC-Store) and data processed on HPC hosts are transferred to ASC bulk storage where image and mesh files are organised in folders by batch, then barcode number, then subfolders for each image file type, including RGB images, IR images, and LiDAR (Light Detection and Ranging Sensors, and their related meshes. Bag creation is carried out via an allocated ASC HPC job. The metadata required for a DAP publication is created and the bag transferred to the DAP staging area via SFTP (SSH File Transfer Protocol). After publication of the DAP collection, the data from PlantScan™ for the given project becomes discoverable via DAP. In addition, experiment reports, published papers, and sensor configurations can either be made accessible via a DAP collection's "related materials" links, other metadata fields, or within the collection's data (e.g. bag).

4.3 Heterogeneous data streams

PlantScan™ is a medium throughput high resolution phenotyping platform, which brings together a number of imaging sensors–light detection and ranging, far-infrared imaging, and multi-wavelength imaging–to non-invasively measure plant growth and function using in-silico approaches. Raw data is captured with its contextual information (e.g. system configuration, time of acquisition, batch number and project) and is stored in a purpose-built database as the data is being generated. The various data streams are collated and used to produce full 3D representation of each plant with overlaid spectral information. The metadata collected during image acquisition are necessary inputs for the computer vision techniques which are used to create the 3D representation of the plant. The 3D meshes are then automatically segmented in order to semantically identify the different parts of the plants [10]. A longitudinal 3D matching pipeline for plant mesh parts is then used to evaluate temporal changes at the whole plant and/or organ level.

4.4 Metadata

Each acquisition on PlantScan™ includes metadata (in addition to the raw data streams), such as plant genus and species, project and experiment metadata, a unique identifier for each image (Globally Unique Identifier), imaging angle, environmental temperature of the imaging chamber, location of optical and colour calibration datasets for each acquisition run, and LiDAR calibration files. The metadata associated with each acquisition is automatically generated when setting up the configuration on the platform. This information is paramount to validate and process the raw image data, and for the post-processing phases.

4.5 Data volume

Digitisation systems such as PlantScan™ generate huge amounts of data including raw image data, registration metadata, sensor configurations and plant metadata. For example, PlantScan™ generates around 500GB of raw image data, representing in excess of 200,000 database records, per day. Sufficient storage space (usually at remote locations) and fast network transfer rates are thus necessary to facilitate data movement for processing using high performance computers (HPC). Because an RDF database structure is not suitable for handling large data sets of images, it is necessary to package the raw information into elementary units with permanent addresses which could be retrieved using PODD. The CSIRO DAP [2] and ASC storage and compute facilities [9] are key resources used by PlantScan™ to process and store bulk data.

5 Semantic integration

The PODD ontology enables plant phenomics researchers to link from mesh results to the raw data that they were generated from. It also allows researchers to link from both mesh results and their recorded conclusions to shared phenomics ontologies which describe specific features of the plants. When used together, this enables scientists to trace the provenance of their results and conclusions based on well known concepts in phenomics ontologies.

Subsets of phenomics ontologies such as the Plant Ontology and the Crop Ontology were mapped into PODD by adding OWL constraints. These constraints enable PODD to verify that the use of classes and properties from these ontologies was consistent with the PODD ontology. For example, the Crop Ontology contains a class defining soil as “Sandy Loam”, giving it the identifier “0000104”. This was mapped into PODD to define a particular soil sample as being Sandy Loam using the triple: *poddSampleSandyLoamSoil a cropOntology* : _0000104.

6 Semantic publication

PODD provides a secure mechanism for publishing both human and machine readable descriptions of scientific experiments. It utilises the well-known DOI mechanism for publishing raw data files using DAP, and uses HTTP URIs to publish experiments using the PODD web interface.

Scientific journals increasingly require the data and provenance for articles to be available in a machine readable format. The DOI registrar that DAP uses, DataCite [11], was setup to provide unique identifiers for data items that can be attached to publications, which in turn may have their own DOIs.

By providing machine readable descriptions of scientific experiments, including semantic references to shared ontologies where possible, PODD enables the output from PlantScan™ to be interpreted and extended by others. The use of PODD URIs in other RDF documents enables scientists to extend the initial work using the Linked Data paradigm [12].

7 Conclusion

This paper described how the Phenomics Ontology Driven Data repository integrates with the PlantScan™ platform and CSIRO Data Access Portal to manage the complex workflows at the High Resolution Plant Phenomics Centre. This workflow keeps track of phenomics data, metadata and experimental processes and also provides a secure mechanism to share and publish scientific experiments in both human and machine readable formats.

References

1. Li, Y.F., Kennedy, G., Davies, F., Hunter, J.: PODD: An ontology-driven data repository for collaborative phenomics research. In Chowdhury, G., Koo, C., Hunter, J., eds.: The Role of Digital Libraries in a Time of Global Change. Volume 6102 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2010) 179–188
2. CSIRO IM&T: CSIRO data access portal. http://data.csiro.au
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature Genet. **25** (2000) 25–29
4. Avraham, S., Tung, C.W., Ilic, K., Jaiswal, P., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Zapata, F., Ware, D.: The plant ontology database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Research **36**(suppl 1) (2008) D449–D454
5. Fürber, C., Hepp, M.: Using sparql and spin for data quality management on the semantic web. In Abramowicz, W., Tolksdorf, R., eds.: Business Information Systems. Volume 47 of Lecture Notes in Business Information Processing. Springer Berlin Heidelberg (2010) 35–46
6. Kunze, J., Littman, J., Madden, L.: The bagit file packaging format (v0.97) (April 15 2011)
7. Summers, E.: Bagit python software. https://github.com/edsu/bagit
8. Library of Congress: Bagit java software. http://sourceforge.net/projects/loc-xferutils/files/loc-bagger/
9. CSIRO IM&T: CSIRO advanced scientific computing. https://wiki.csiro.au/display/ASC
10. Paproki, A., Sirault, X., Berry, S., Furbank, R., Fripp, J.: A novel mesh processing based technique for 3d plant analysis. BMC Plant Biology **12**(1) (2012) 63
11. Brase, J.: Datacite - a global registration agency for research data. In: Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO '09. Fourth International Conference on. (2009) 257–261
12. Berners-Lee, T. http://www.w3.org/DesignIssues/LinkedData.html (2006)

Lightweight Ontology-Based Tools for Managing Observational Data

Shawn Bowers, Riley Englin, Carlos Fonseca, Paul Jewell, Lauren Joplin, Patrick Mosca, Tyler Pacheco, Jacob Troxel, Tyler Weeks

Deptartment of Computer Science, Gonzaga University, Spokane, WA, USA

Abstract. We describe recent ontology and annotation editing capabilities to a specialized data management system for observational data. The system supports observations and measurements explicitly, allowing users to upload observational data sets as well as semantically describe and query data sets using formal OWL-DL ontologies. Recent extensions allow users to extend observational ontologies with domain-specific terms as well as provide detailed semantic annotations using a "markdown"-based approach. In addition, we describe a new implementation of the system using standard semantic web technologies for managing OWL-DL ontologies and RDF triples. Our approach supports a wide variety of observational data, and is especially targeted at helping scientists manage heterogeneous biodiversity and ecological data by allowing access to data through a common and generic observations and measurements data model.

1 Introduction

Performing an ecological analysis to study phenomena across geographic, temporal, or biological scales typically requires access to a variety of existing (already collected) observational data sets. A major challenge when performing such an analysis is understanding and reconciling the structural and semantic differences among data sets. In particular, data sets often differ in the number of attributes, the names of similar attributes, the relationships implied between attributes, and the coding conventions used for representing information within data sets. These differences not only make discovering relevant data difficult, but also requires researchers to spend considerable time interpreting and integrating potential data sets for use within any particular analysis. We aim to help address these challenges by providing a suite of lightweight, ontology-based tools that allow researchers to semantically describe, access, and analyze heterogeneous data sets (either their own, or those collected for use in research studies). In this paper, we describe tools that have recently been developed within the ObsDB system [5], which provides data management support built on top of a generic ontology model for formally representing observations and measurements [6]. Within ObsDB, data sets are viewed as semantically described collections of observations. In particular, when data is registered with ObsDB, it is converted automatically into the appropriate observational structure (and represented within the current version of ObsDB as an RDF graph). This approach allows otherwise hard to manage, heterogeneous table structures to be viewed and accessed uniformly as collections of observations and measurements.

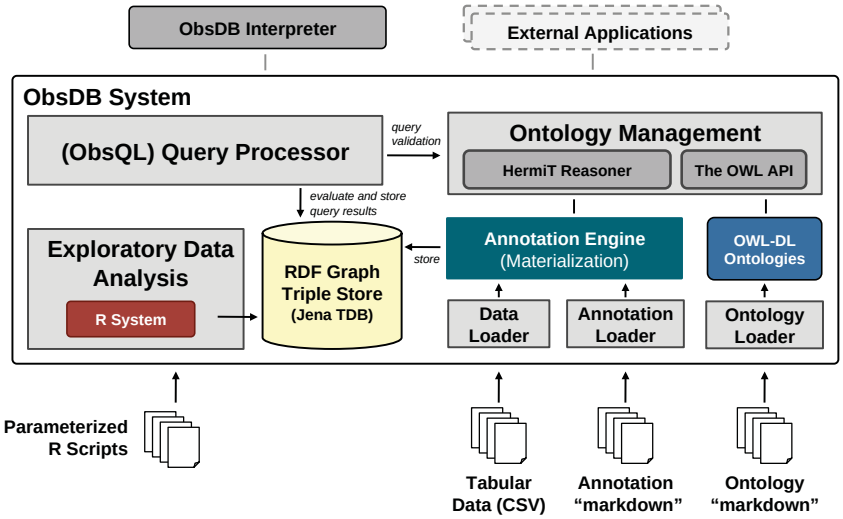

Fig. 1. Overview of the ObsDB system architecture.

The tools we have developed leverage a "markdown"-based approach for defining ontologies, for specifying data-level semantic annotations (from which data is "shredded" into the observational model), and for expressing data discovery queries. The "markdown" is then converted automatically within ObsDB (see Figure 1) into corresponding descriptions in OWL-DL (for representing ontologies), RDF (for storing observation and measurement instances), and SPARQL (for executing discovery queries and filtering collections). ObsDB also allows observations and measurements from multiple data sets to be merged into "virtual" observation collections (RDF graphs), which can be saved and further accessed and queried. Ontologies within ObsDB allow observations and measurements to carry rich semantic descriptions, including the types of entities observed, the characteristics of entities measured, the context in which observations were made, and detailed measurement standards (units) for facilitating unit conversion. ObsDB also supports an expressive query language for selecting data sets and filtering data by observation and measurement types. For instance, users can specify queries to find all data sets that contain specific measurements of entities (e.g., diameter and height measurements of trees), relationships and constraints among entities (e.g., the length of branches on trees of a minimum height and at specific elevations), and the use of desired measurement standards (e.g., in meters). Similar queries can also be expressed to obtain all observations (either within or across collections) matching such criteria, where unit conversions are automatically applied as needed. Through integration with the R system [1], analytical scripts can also be called from within ObsDB to perform a variety of exploratory analyses over observation collections.

[1] `http://www.r-project.org`

In this paper, we extend our prior work on ObsDB [5,8] by describing recent extensions to the system, focusing in particular on new support for ontology modeling, annotation, and querying, and its implementation over underlying semantic web technologies. We demonstrate our "markdown"-based approach using examples drawn from real-world ecological data, and also describe our ongoing and future work on further extending ObsDB with the goal of helping researchers more effectively manage heterogeneous observational data.

Figure 1 shows the main architectural components of ObsDB. The ObsDB system is implemented in Java and can be used from within other (external) applications (via API calls) or by using the ObsDB interpreter. ObsDB manages user loaded data set files, semantic annotation files, and ontology files. Ontologies are converted to OWL-DL files by ObsDB and are stored and managed using the OWL API[2]. Annotation files can be applied to data sets to produce a "materialized" set of RDF triples (an RDF Graph). All RDF data is stored within ObsDB using the Jena triple store[3] technology. Users can query RDF Graphs using the ObsDB query processor, which converts high-level queries expressed in ObsQL (the query language of ObsDB) into corresponding SPARQL queries. As part of the query evaluation process, ObsDB uses the HermiT OWL-DL reasoner for query expansion (which is also used to verify semantic annotations are semantically consistent). Finally, R scripts can be defined and registered with ObsDB to perform statistical and analytical operations over RDF Graphs stored within ObsDB.

The rest of this paper describes these features in more detail. Section 2 describes the underlying observations ontology employed by ObsDB and its newly supported ontology "markdown" approach. Section 3 describes the new semantic annotation approach employed by ObsDB. Section 4 briefly describes ObsQL and its new implementation in ObsDB. Finally, Section 5 concludes by describing related work and our future directions for ObsDB.

2 Ontology Creation and Management

The ObsDB system is built on a recent version of the Extensible Observation Ontology (OBOE) [6][4]. The OBOE model is implemented in OWL-DL and is compatible with the O&M ISO standard developed by the Open Geospatial Consortium (OGC) [1]. Figure 2 shows the top-level classes and properties supported by OBOE (the primary "OBOE core" classes). An *observation* is made of an *entity* (e.g., biological organisms, geographic locations, or environmental features, etc.) and primarily serves to group a set of measurements together to form a single observation event. A *measurement* assigns a value to a *characteristic* of the observed entity (e.g., the height of a tree), where a value is represented through a special class (similar to the notion of value partitions in [15]). Measurements also include *standards* (e.g., units) for relating values across measurements, and can specify additional information including collection protocols,

[2] `http://owlapi.sourceforge.net/`

[3] `http://jena.apache.org/documentation/tdb/`

[4] See `https://code.ecoinformatics.org/code/semtools/trunk/dev/oboe/oboe.1.1rc1/oboe-core.owl`

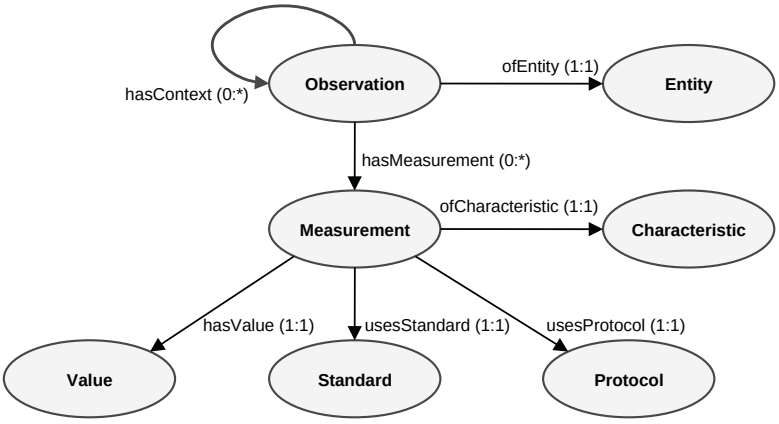

Fig. 2. Main concepts and properties in the ObsDB observations ontology (OBOE).

methods, precision, and accuracy (not all of which are shown in Figure 2). An observation (event) can occur within the *context* of zero or more other observations, e.g., an observation of a tree specimen may have been made within a specific geographic location, and the geographic location provides important information for interpreting and comparing tree measurements. In this case, by establishing a context relationship between the tree and location observations, the measured values of the location are assumed to be constant with respect to the measurements of the tree. Context forms a *transitive* relationship among observations. A key feature of the model is its ability for users to assert properties of entities (as measurement characteristics or contextual relationships) without requiring these properties to be interpreted as inherently (or always) true of the entity. Depending on the context an entity was observed in, its properties may take on different values. For instance, the diameter of a tree changes over time, and the diameter value often depends on the protocol used to obtain the measurement. The observation and measurement structure of Figure 2 allows RDF-style assertions about entities while allowing for properties to be contextualized (i.e., the same entity can have different values for a characteristic under different contexts), which is a crucial feature for modeling scientific data [6]. The primary differences between O&M and OBOE are that (1) OBOE was designed to explicitly be represented in OWL-DL; and (2) OBOE treats an observation (event) as a collection of measurements, allowing observations to be defined within a context hierarchy (which implicitly applies to an observation's associated measurements) as opposed to O&M which requires each measurement's context to be stated explicitly.

Figure 3 shows the main classes and properties defined in OBOE for representing measurement standards, including units of measure. Every measurement unit is associated with a measurement characteristic (e.g., length, mass, time, area, volume, etc.) and the set of units are divided into four subclasses. A *base unit* represents a unit that cannot be naturally divided into smaller units. Examples include meter, gram, second, kelvin, and so on. A *prefixed unit* applies a prefix (represented as a literal value) to a

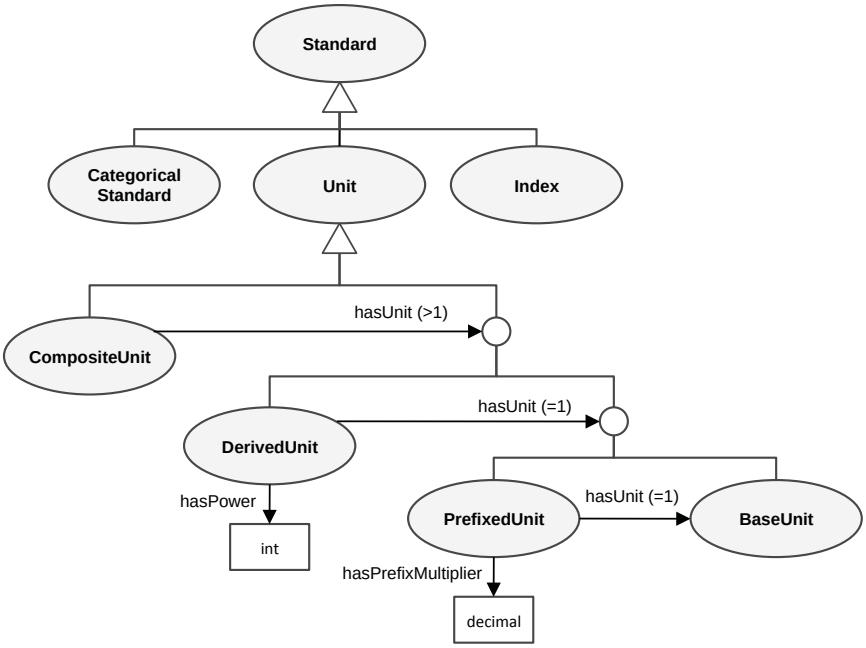

Fig. 3. The basic class hierarchy for describing measurement units (OBOE).

base unit. Examples include kilogram (with base unit gram and multiplier 1000) and centimeter (with base unit meter and multiplier 0.01). A *derived unit* assigns a power (other than 1) to either a prefixed or base unit. Examples include meter squared (m^2), hertz (s^-1), and microliter (mm^3). A *composite unit* combines 2 or more derived, prefixed, or base units. Examples include meter per second (which is composed of meter and a derived per second unit, i.e., $m \times s^{-1}$) as well as "dimensionless" units like gram per gram (g/g) in which retaining the original units is often needed for interpreting and integrating data.

Users of ObsDB can create their own ontologies that extend OBOE with domain-specific terms when annotating data. ObsDB supports a lightweight "markdown" syntax for describing terms, which was developed specifically to support OBOE modeling constructs and common modeling patterns. The advantage of having a lightweight syntax is that it allows non-expert users in OWL-DL to create and edit terms as needed, without needing to learn DL syntax (required, e.g., in Protege[5] or if editing OWL-DL/RDF syntax directly). Ontologies expressed in the lightweight syntax are automatically converted to the corresponding OWL-DL representation by ObsDB (see Figure 1). The lightweight syntax is based on YAML[6] and provides support for the following tasks:

(i). Importing domain-specific OBOE ontologies for extension;

[5] `http://protege.stanford.edu/`
[6] `http://www.yaml.org/`

(ii). Creating entity, characteristic, and protocol hierarchies;

(iii). Defining characteristic qualifiers (e.g., to specify an "average" length characteristic where "average" denotes the qualifier);

(iv). Creating base, prefixed, derived, and composite units;

(v). Specifying unit conversions; and

(vi). Defining categorical measurement standards.

For instance, the following example establishes a simple class hierarchy using the ObsDB lightweight ontology syntax:

```
!Entity
   name: "Organism"
   childClass:
      !Entity
      name: "Tree"
      # basic types of trees
      childClass:
         !Entity
         name: "DominantTree"
         comment: "A tree that extends above surrounding ...
      childClass:
         !Entity
         name: "OvertoppedTree"
         comment: "A tree that cannot sufficiently extend its crown ...
         equivalentClass: "SuppressedTree"
      ...
      # different species of trees
      childClass:
         !Entity
         name: "DouglasFir"
         equivalentClass: "Pseudotsuga_menziesii"
         comment: "See Garrison et al., 1972"
      ...
```

In this example, a tree class is defined as a subclass of a generic organism class. The tree class is also defined with three subclasses (dominant, overtopped, and Douglas fir). Each class has a name and an optional comment. In addition, equivalent classes (i.e., synonyms) can be specified as in the case of an overtopped tree (defined as being equivalent to a suppressed tree) and with Douglas fir (where in addition to the common name the taxonomic name is also given). The following defines an example physical characteristic.

```
import char: "http://code.ecoinformatics.org/.../oboe-characteristics.owl"
!PhysicalCharacteristic
   name: "DiameterAtBreastHeight"
   parentClass: "char:Diameter"
```

Here diameter at breast height (DBH) is defined as a subclass of the diameter class, which is imported from another ontology (as given by the `import` statement). The following example defines a simple base unit, composite unit, and unit conversion.

```
!BaseUnit
   name: "Meter"
   characteristic: "oboe:Length"
!CompositeUnit
   name: "MeterPerSecond"
   characteristic: "oboe:Speed"
   allUnits:
      - "Meter"
      - !Derived
```

```
            baseUnit: "Second"
            power: -1
!UnitConversion
   name: "FootToMeter"
   source: "Foot"
   target: "Meter"
   multiplier: 0.3048
   offset: 0
```

In this example, the composite unit is defined over the base unit meter and a derived unit defined "on the fly" (i.e., without providing a specific name to the unit). Finally, the following illustrates a simple categorical standard definition.

```
!CategoricalStandard
   name: "TreeGrowthVigorStandard"
   comment: "Standard values for good, fair, and poor tree growth vigor"
   values: "TreeGrowthVigorValue" {"good_tree_growth_vigor",
              "fair_tree_growth_vigor", "poor_tree_growth_vigor"}
```

In this case, we are defining a value partition (as in [15]) consisting of three values representing good, fair, and poor tree growth.

After starting the ObsDB interpreter, users can load ontology files using the `load onto` command. When loading an ontology file, a namespace prefix and URI is also assigned to the ontology for use within ObsDB. For instance, the following shows the result of starting ObsDB and loading the "ont1.yml" ontology file:

```
ObsDB v1.0
Type 'help' for a list of commands. Type 'quit' to quit ObsDB.
> import onto 'ont1.yml' as 'ont1' using 'http://obsdb.org/ont1'
Ontology created
Ontology loaded
```

In this case, we are assigning the ontology the namespace prefix "ont1" and the URI "http://obsdb.org/ont1". Once loaded, the ontology can be accessed via the namespace. Ontologies can also be updated using the ObsDB `update onto` command, which allows ontologies to be modified without having to remove (or `drop`) an ontology and then load the updated version.

In general, we have found that using a lightweight approach such as this has a number of benefits for specifying OBOE extensions and for annotating data. In particular, the approach allows new ontology terms and entire ontologies to be quickly and easily created by simply opening and editing a text file, and the high-level syntax supports otherwise complex description-logic definitions without requiring users to be experts in description logic (which is often the case in Protege's OWL-DL editor). The latter is especially an issue in ontologies like OBOE that leverage description logic constraints that must be maintained, e.g., via modeling patterns such as value partitions and various class and property restrictions. Once a user loads an ontology into ObsDB, the system automatically performs syntax and semantic validation (e.g., checking for inconsistencies). Together with the lightweight text-based syntax, this allows for rapid editing, loading, and validation of OBOE ontology extensions.

3 Observational Data Sets and Semantic Annotations

Semantic annotations in ObsDB define how to translate a tabular data set into a corresponding collection of semantically relevant observations and measurements. Semantic

STAND	PLOT	TAG	SPP	YEAR	DBH	CANCLASS	VIGOR
B388	1	3319	PSME	1999	22.5	C	1
B388	1	3320	PSME	1999	16	I	1
B388	2	3336	PSME	1999	33	D	1
B388	2	3339	CACH	1999	5.8	S	1
B646	1	1817	PSME	1999	22	C	1
B646	1	1815	CACH	1999	5.7	I	1
B684	2	2207	ALRU	1999	19.9	C	1
...	...	...	...	...	...	...	...

Fig. 4. Example data set consisting of tree (allometry) observations and measurements.

annotations are defined using *semantic templates* [8] that specify the observation and measurement types (and their various relationships) for the data set. The observations and measurements given by each template are automatically filled in (to create observation and measurement instances) based on user-defined mappings from data set attributes to measurement types. For instance, consider the example data set in Figure 4. This data set[7] consists of eight attributes and approximately three thousand rows of data (only six of the rows are shown in the figure). The first two attributes specify contextual information concerning the stand and plot where the tree was observed. We can annotate these attributes using the annotation "markdown" syntax supported by ObsDB as follows. For instance, the attribute denoting the stand is annotated by the semantic template:

```
import ont1: 'http://obsdb.org/ont1'
observation 'StandObs':
   entity: 'ont1:Stand'
   measurement:
      characteristic: 'obs:Name'
      value: '$STAND'
   entityKey: '$STAND'
```

which creates an observation individual of a stand entity for each unique value of the STAND attribute in the data set (thus, for stand B388, B646, and B684 in Figure 4). The measurement in this case is simply the name of the stand, which is taken directly from the attribute values. The entityKey field of the template specifies that each unique value should generate a new observation (as opposed to each row, regardless of the STAND value, generating a new observation). In this example, we also import a domain-specific ontology and assign a namespace prefix to be used to refer to corresponding classes within the annotation file. Similar to the stand template, the following template can be used to annotate the plot information in the data set.

```
observation 'PlotObs':
   entity: 'ont1:Plot'
   measurement:
      characteristic: 'obs:Name'
      value: '$PLOT'
   context: 'StandObs'q
   entityKey: '$PLOT' within 'StandObs'
```

[7] Based on one of the many data sets available on the H.J. Andrews Experimental Forest LTER site (`http://andrewsforest.oregonstate.edu/`).

Here the plot is nested within the stand, and so each plot observation has as context the corresponding stand observation. In this data set, the names of plots across stands are not unique (e.g., stand B288 contains a plot 1 as does stand B646). This information is denoted using the "within" keyword. Although not included here, additional measurements can also be added to the template, e.g., the area of the plot (which in this case could be specified as a constant value assuming all plots are of the same area). The year attribute would be annotated similarly to the stand as follows.

```
observation 'YearObs':
   entity: 'ont1:TimePeriod'
   measurement:
      characteristic: 'ont1:Year'
      value: '$YEAR'
   entityKey: '$YEAR'
```

The remaining attributes would be annotated via a single tree observation template:

```
observation 'TreeObs':
   entity: match '$CANCLASS' with
      'D' => 'ont1:DominantTree'
      'S' => 'ont1:SuppressedTree'
      ...
   entity: match '$SPP' with
      'PSME' => 'ont1:Pseudotsuga_menziesii'
      'CACH' => 'ont1:Castanopsis_chrysophylla'
      'ALRU' => 'ont1:Alnus_rubra'
      ...
   measurement:
      characteristic: 'obs:Name'
      value: '$TAG'
   measurement:
      characteristic: 'ont1:DiameterAtBreastHeight'
      standard: 'ont1:Meter'
      value: '$DBH'
   measurement:
      characteristic: 'ont1:Vigor'
      standard: 'ont1:TreeGrowthVigorStandard'
      value: match '$VIGOR' with
         '1' => individual: 'ont1:good_tree_growth_vigor'
         '2' => individual: 'ont1:fair_tree_growth_vigor'
         '3' => individual: 'ont1:poor_tree_growth_vigor'
   context: 'PlotObs', 'YearObs'
   entityKey: '$TAG'
```

In this example, the entity is defined by two separate attributes: the CANCLASS attribute specifies the canopy class the tree belongs to; and the SPP attribute specifies the tree species. In both cases, attribute values are mapped to ontology classes (i.e., a value such as "PSME" denotes a specific species type). Each tree observation also has three associated measurements: the (tag) name of the tree, the diameter at breast height (measured in meters); and the growth vigor. For the tree's vigor, each value in the dataset is mapped to a specific OWL-DL individual value (as opposed to a class) that is defined in the corresponding vigor standard of the ontology example of Section 2. Finally, each tree observation is made within the context of its corresponding plot and the year in which the observation was made.

Data sets and annotations are loaded independently in ObsDB. The `import table` command is used to register a CSV file denoting a data set with ObsDB. For example, the following command can be used to load the example table of Figure 4.

```
> import table 'table1.csv' as 'table1'
File copied to /data/tables/ directory.
File loaded.
```

Annotations are loaded into ObsDB using the `import annotation` command:

```
> import annotation 'annot1.txt' as 'annot1'
File copied to /data/annotations/ directory.
File loaded.
```

Once loaded, annotations can be applied to tables to generate an RDF graph of corresponding observation and measurement instances using the ObsDB `apply` command:

```
> apply 'annot1' to 'table1' as 'coll1' using 'http://obsdb.org/coll1'
***Generating Data From Files***
Annotation: data/annotations/annot1.oal
Table: data/tables/table1.csv
Output Triples: data/graphs/coll1.ttl
```

Here "coll1" is used to name the resulting named RDF graph, which is given the corresponding URI. ObsDB performs the following steps when applying an annotation to a table: (1) it verifies the annotation and data file are both syntactically correct; (2) it verifies the imported ontologies exist and that they have been loaded into ObsDB; (3) it generates the corresponding observations and measurements (i.e., by *materializing* the semantic annotation templates); and (4) it uses the HermiT OWL-DL reasoner to add inferred axioms (based on ontology definitions and constraints) to the RDF Graph, and ensures the resulting graph is consistent. Adding inferred axioms is performed to support query expansion within ObsDB. The resulting RDF Graph is stored within ObsDB and can then be further accessed and queried. [8] describes an earlier version of the materialization algorithm used by ObsDB. The approach used in the current version of ObsDB extends this work by supporting more complex value matching annotation primitives (as shown above, e.g., with the canopy class and vigor measurements) as well as by materializing data sets to named RDF Graphs as opposed to an underlying relational database representation.

4 Data Discovery and Analysis

Once data sets are semantically annotated and converted into their corresponding RDF graphs, they can be accessed directly from within ObsDB. There are three main ways to access data sets: (1) using the `find` command to issue data discovery queries to locate observation collections (RDF graphs of observations and measurements); (2) using the `query` command to select observations within or across data sets, the results of which can be viewed or used to create new observation collections that subset existing collections or combine multiple existing collections; or (3) using the `exec` command to apply statistical and analytical functions to query results (using built-in aggregation operators or by calling external R scripts).

Each of the above ways to access observation collections are expressed in ObsDB using the high-level query language ObsQL [6]. ObsQL queries are similar in spirit to XPath queries for XML in that ObsQL is designed to provide a simple syntax for expressing common data discovery and subsetting operations. For example, the following ObsQL find expression can be used to locate all observation collections that contain observations of trees:

```
> find ont1:Tree []
Matching Graphs
---------------
coll1
```

This example returns the set of matching observation collections (i.e., RDF graphs), which in this case consists of the "coll1" example of Section 3. All ObsQL find and query expressions are rewritten by ObsDB into corresponding SPARQL queries. For instance, the above ObsQL expression is converted by ObsDB into the SPARQL ASK query:

```
PREFIX obs: <https://code.ecoinformatics.org/.../oboe-core.owl#>
PREFIX char: <https://code.ecoinformatics.org/.../oboe-characteristics.owl#>
PREFIX ont1: <http://obsdb.org/ont1#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
...
ASK {?temporaryObservationVariable0 obs:ofEntity ?temporaryVariable0 .
?temporaryVariable0 rdf:type ont1:Tree .}
```

As another example, we can refine the search above to look for only those collections that have a specific type of tree with a specific type of measurement:

```
> find ont1:DouglasFir [ont1:DiameterAtBreastHeight]
Matching Graphs
---------------
coll1
```

ObsQL also supports context constraints using the -> operator. For example, the following expressions locates collections that contain observations of Douglas Fir trees made within named plots:

```
> find ont1:DouglasFir [ont1:DiameterAtBreastHeight] -> ont1:Plot [obs:Name]
Matching Graphs
---------------
coll1
```

While the above expression is relatively straightforward to express in ObsQL, the corresponding SPARQL query is considerably more verbose:

```
...
ASK {?tempObsVar0 obs:ofEntity ?tempVar0 .
     ?tempVar0 rdf:type ont1:DouglasFir .
     ?tempObsVar0 obs:hasMeasurement ?tempMeasVar0 .
     ?tempMeasVar0 obs:ofCharacteristic ?tempPlaceholder0 .
     ?tempPlaceholder0 rdf:type ont1:DiameterAtBreastHeight .
     ?tempMeasVar0 obs:hasValue ?tempPlaceholder1 .
     ?tempPlaceholder1 obs:hasCode ?tempVar1Code .
     ?tempObsVar1 obs:ofEntity ?tempVar2 .
     ?tempVar2 rdf:type ont1:Plot .
     ?tempObsVar1 obs:hasMeasurement ?tempMeasVar1 .
     ?tempMeasVar1 obs:ofCharacteristic ?tempPlaceholder2 .
     ?tempPlaceholder2 rdf:type obs:Name .
     ?tempMeasVar1 obs:hasValue ?tempPlaceholder3 .
     ?tempPlaceholder3 obs:hasCode ?tempVar3Code .
     ?tempObsVar0 obs:hasContext ?tempObsVar1 .}
```

We note that while ObsQL expressions are generally more concise and easier to specify than their corresponding SPARQL queries (in part, because ObsQL is tailored specifically to supporting queries over OBOE models), only a subset of SPARQL can be expressed in ObsQL. In addition to the above example, it is also possible to specify multiple contexts for an observation, for example:

```
> find ont1:IntermediateTree [] -> (ont1:Stand [], ont1:TimePeriod [])
Matching Graphs
---------------
coll1
```

finds all collections with an observation of the given tree type within the context of both a stand and a time period. Note that here the stand is an indirect context for the corresponding tree since the tree has a plot as context, and the plot has the stand as context. It is also possible to query for observations and measurements within a collection. For example, the following query returns the diameter values of all codominant trees within coll1:

```
> query ont1:CodominantTree [ont1:DiameterAtBreastHeight $d] in coll1
---------------------------------
| temporaryVariable0 | d        |
=================================
| :ID1004            | "10.3"   |
| :ID1017            | "13.1"   |
| :ID10286           | "48.9"   |
| :ID10299           | "46"     |
...
```

Here, $\$d$ is a "place holder" variable for specifying output values. Note that although not shown here, multiple place holder variables can be given per query. Also, removing the "in" clause above will result in ObsDB querying all collections for matching observations.

Basic computations can also be performed on data when using ObsQL. For instance, when querying it is possible to select a specific unit from which ObsDB will apply appropriate unit conversions. For example, in this query:

```
> query ont1:CodominantTree [ont1:DiameterAtBreastHeight $d ont1:Foot] in coll1
---------------------------------------------
| temporaryVariable0 | d                    |
=============================================
| :ID1004            | "33.79265091863517"  |
| :ID1017            | "42.979002624671914" |
| :ID10286           | "160.43307086614172" |
| :ID10299           | "150.91863517060366" |
...
```

ObsDB uses the foot-to-meter unit conversion of Section 2 to convert the diameters in coll1 from meters to feet (since conversions are invertible). As another example, ObsDB can also perform statistical summaries of query results. For instance, the following query computes for each plot the average Douglas Fir tree diameter (in meters):

```
> exec avg $d by $p in coll1 where
     ont1:DouglasFir [ont1:DiameterAtBreastHeight $d ont1:Meter]
     -> ont1:Plot $p
-----------------------
| ?p       | mean     |
=======================
| :ID908   | 13.4     |
| :ID10203 | 59.51429 |
| :ID7     | 24.6     |
| :ID10837 | 155.4143 |
...
```

ObsDB supports the standard aggregate operations supported by SPARQL including average, mean, count, max, min, median, range, and standard deviation. Finally, custom R scripts can be used from within ObsDB. Each R script must contain a comment header denoting the name of the operation (for use within an `exec` command) and the variables that will be passed into the script (the script inputs). As a simple example, the following commented R script can be used to draw a basic histogram from within ObsDB.

```
#name: hist
#argument: $x A vector of x-axis variables
x=$x
hist(x)
```

Once defined, this script can be called from within ObsDB as follows.

```
> exec hist $d in coll1 where
    ont1:CodominantTree [ont1:DiameterAtBreastHeight $d > 75 ont1:Meter]
```

The result of this command generates the histogram shown in Figure 5, showing the distribution of codominant tree diameters greater than 75 m in the underlying data set. This example also demonstrates an ObsQL query that performs a logical comparison on measurement values.

In general, ObsQL is designed to provide scientists with the ability to search for relevant data sets based on domain-specific ontology classes as well as perform basic exploratory analyses through the subselection of relevant observations and measurements of data sets, by applying aggregate operations, and by applying simple R analysis scripts. Although not shown in the examples above, ObsDB also allows the results of queries to be stored in new observation collections using the `as` keyword. This provides a basic form of data integration, in which observations from multiple data sets can be combined into a single collection, without having to perform similar operations directly on structurally heterogeneous tabular data sets.

5 Related Work and Future Directions

This paper described extensions to our prior work on semantic annotation and providing access to observational data through the OBOE model [6]. An early version of ObsDB was presented in [5], which did not directly support ontology editing and semantic annotations. Instead, the approach assumed that data was already "materialized" into RDF triples, which could then be loaded into the system. Similarly, ontologies were assumed to be defined outside of the system and accessible through resolvable URIs. The early version of ObsDB also employed a relational database system for storing and querying observational data, which lead to a number of performance issues as well as limiting its interoperability with other semantic web technologies. In [8] we described an approach for supporting semantic annotation templates. This work also relied on observations and measurements being stored using a relational database system. [8] also gives a formal specification of annotations along with alternative implementation strategies (in the spirit of classical data integration view-based approaches, e.g., [12,11]). Taken together, we extend our prior work by providing: an ontology editing "markdown" language designed specifically for OBOE; additional annotation constructs (for value mappings);

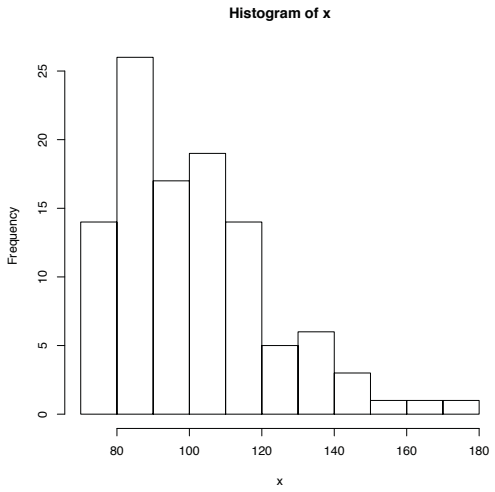

Fig. 5. Example histogram produced via an R script run in ObsDB.

a new ObsDB implementation over a popular RDF triple store system (Jena TDB); automated unit conversion; and extended query capabilities (e.g., value comparisons, grouping operations, and a wide range of aggregate operations).

A number of systems have recently been developed that leverage observation-based conceptual and ontology models. Some examples include [13] which describes and compares four implementations of the OGC's sensor observations service [2] for accessing and querying real-time sensor data (and which relies on O&M for representing observations), [10] which describes a system for managing resources based on hydrological observations (focusing on supporting large-scale observatory networks), and [16] which describes an implementation of a water quality portal that integrates a number of data sets via observational ontologies. ObsDB largely differs from these approaches by providing a personal data management system (as opposed to targeted applications over an observational model), with support for ontology-based data annotation, ontology editing, and various forms of exploratory query support. Annotations have been studied in various forms in the database literature (e.g., [9,4]), and as mentioned previously, our approach is similar to the more general use of views in data integration. Finally, the need for uniform mechanisms to describe observational data has led to many proposals for observational data models and ontologies (e.g., [14,3,7]). ObsDB is largely complementary to these efforts by providing a framework for managing observational data according to a generic observational model (based on OBOE) that supports the use of domain-specific ontologies, and a high-level query language for discovering and accessing observations (within and across datasets).

Our ongoing and future work on ObsDB is focused on further extending support for exploratory analysis of observational data via the R system. We are also interested in developing tools within ObsDB to support comparing data sets based on their annotations. For instance, given two observation collections, we would like to determine

how closely they "match" in terms of observation and measurement types and to automatically create mappings and transformations to unify the collections into a single integrated annotation template (for further analysis). We are also interested in developing support in R to access observations stored in ObsDB, e.g., to be able to programmatically load one or more observation collections through R calls to perform more sophisticated analyses (within an R script).

Acknowledgements

This work supported in part through NSF grants IIS-1118088, DBI-0743429, and DBI-0753144.

References

1. OGC: Observations and measurements encoding standard (O&M): http://www.opengeospatial.org/standards/om
2. OGC: Sensor observation service (SOS): http://www.opengeospatial.org/standards/sos
3. Semantic Web for Earth and Environmental Terminology (SWEET), http://sweet.jpl.nasa.gov/sweet/
4. An, Y., Mylopoulos, J., Borgida, A.: Building semantic mappings from databases to ontologies. In: AAAI (2006)
5. Bowers, S., Kudo, J., Cao, H., Schildhauer, M.P.: Obsdb: A system for uniformly storing and querying heterogeneous observational data. In: eScience. pp. 261–268 (2010)
6. Bowers, S., Madin, J.S., Schildhauer, M.P.: A conceptual modeling framework for expressing observational data semantics. In: ER. pp. 41–54 (2008)
7. C. Mungall, *et al.*: Integrating phenotype ontologies across multiple species. Genome Biology 11(R2) (2010)
8. Cao, H., Bowers, S., Schildhauer, M.P.: Approaches for semantically annotating and discovering scientific observational data. In: International Confere on Database and Expert Systems Applications (DEXA). pp. 526–541 (2011)
9. Geerts, F., Kementsietsidis, A., Milano, D.: Mondrian: Annotating and querying databases through colors and blocks. In: ICDE. p. 82 (2006)
10. Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I.: Components of an environmental observatory information system. Computers & Geosciences 37(2), 207–218 (2011)
11. Kolaitis, P.G.: Schema mappings, data exchange, and metadata management. In: PODS (2005)
12. Lenzerini, M.: Data integration: a theoretical perspective. In: PODS. pp. 233–246 (2002)
13. McFerren, G., Hohls, D., Fleming, G.: Evaluating sensor observation service implementations. In: IEEE International Geoscience & Remote Sensing Symposium (IGARSS). pp. 363–366 (2009)
14. P. Fox, *et al.*: Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience. Computers & Geosciences 35(4), 724–738 (2009)
15. Rector, A.L., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., Wroe, C.: Owl pizzas: Practical experience of teaching owl-dl: Common errors & common patterns. In: EKAW. pp. 63–81 (2004)
16. Wang, P., Fu, L., Patton, E.W., McGuinness, D.L., Dein, F.J., Bristol, R.S.: Towards semantically-enabled exploration and analysis of environmental ecosystems. In: eScience. pp. 1–8 (2012)

BirdWatch—Supporting Citizen Scientists for Better Linked Data Quality for Biodiversity Management

Eero Hyvönen, Miika Alonen, Mikko Koho, and Jouni Tuominen

Semantic Computing Research Group (SeCo),
Aalto University and University of Helsinki, Finland
first.last@aalto.fi, http://www.seco.tkk.fi/

Abstract. Observational data about species of public interest, such as birds and butterflies, is often created and collected by volunteered citizen scientists, and used by professionals for managing biodiversity. The education and skills of the citizens participating in the work varies a lot, and the process of making observations is typically not systematic but rather ad hoc. As a result, the quality of the observational data in repositories, such as the Global Biodiversity Information Facility GBIF Data Portal, is often not good, hampering its utilization severely. This paper presents an approach for enhancing data quality in a citizen science setting, and presents a mobile tool BirdWatch for citizen observers, mitigating difficulties in producing high quality Linked Data for biodiversity management.

1 Introduction

Biodiversity [1, 6] management (BM) is based on observations of the nature. Special concerns of the field include changes in our environment that lead to undesired changes in the populations of organisms, such as the spread of harmful invasive alien species or extinction of endangered species. Based on observations and their time series, such changes can be identified in time and necessary measures of nature conservation be initiated.

In many areas of biology, much of the observational data is based on citizen science: the data comes from masses of amateurs observing plants, animals, and other organisms of their interest. In this way it is possible to gather lots of useful data for minimal costs. Such data is systematically collected in databases in many countries and also aggregated by organizations such as GBIF[1] on an international level. Today, the GBIF Data Portal includes nearly 400 million observations in over 10,000 datasets, hosted in a network of servers of ca. 420 nature organizations around the world.

Observing the nature and reporting findings is getting more and more popular, and many of the GBIF datasets are based on the observations of volunteered

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

amateurs. The most active domain of biology here is ornithology. In Finland, for example, there are over 10,000 active birdwatchers[2] reporting their observations to databases, about 0.2% of the whole population. The amateurs are equipped with varying knowledge and skills, and the process of making observations is typically rather self-organizing and ad hoc than systematic. As a result, the *quality* of observations varies a lot in different ways:

1. **Misinterpretations** There are lots of misinterpretations of species in the data, e.g., an arctic tern reported as a common tern.
2. **Uncertainty** The observations and data may be uncertain, which may be difficult to represent in a harmonized way.
3. **Trust** Data from an experienced ornithologist should be more reliable than data from a beginner, but this cannot usually be represented and evaluated.
4. **Incompleteness** The data may be incomplete. For example, values may be missing from records, or data in one dataset lacks certain metadata element values or describe them at a different level of granularity.
5. **Statistical biases** The data is statistically concentrated on certain areas, times, and on certain species of interest to the public. Especially big and beautiful species are frequently reported, as well as early or late observations of migratory species.
6. **Machine Interpretability** Observations are represented in different syntactic ways and often using natural language phrases that may be difficult to interpret by the machine.
7. **Interoperability** Metadata about the observations is represented using different models, and different species lists [12] may be in use in different countries.

This paper presents a solution and an online tool that can be used for supporting citizen scientists in producing better quality observational data for biodiversity management. We argue that at least the following requirements are needed for such a system:

1. **Make use of statistical data** of related observations based on the current spatio-temporal context. If someone is trying to report an observation that is very different from the others made at the same place and time before (e.g., a swallow in winter time in Finland), there is a particularly high risk of misinterpretation. Supporting or refuting the observational data of other observers should be provided at the time and place where a new observation is being considered and reported.
2. **Provide identification support based on species characteristics**. Information about the characteristics of the proposed species and related species that look or sound similar is crucial when identifying species.
3. **Shorten the learning curve and boost motivation**. It may take several years to become a reliable nature observer, say an ornithologist. The system should therefore speed up this process by 1) shortening training time and 2)

[2] http://www.birdlife.fi/

also keeping the observer motivated in continuing in her hobby. Providing statistical [4] and ontological data about the species not only helps the end-user in making the identification right, but also teaches her, so that in the future higher quality observations are possible with less help.

4. **Help in creating interoperable data**. Creating observation records is tedious manual work that also distracts the observer from the main task of making observations. The system should therefore help the end-user in creating the observation data record. The data should also be represented in a machine readable, unambiguous, and interoperable way so that its can be processed later correctly and aggregated with other observations.

This paper presents an approach and an online system, "BirdWatch—Mobile Semantic Service for Birding" addressing these issues. As a methodological and technological basis, Semantic Web[3] and Linked Data [3] are used. A major technical novelty of the BirdWatch system is its ability to use and mix both statistical data, based on observation databases, and ontological a priori knowledge about the application domain, in this case birds and their characteristics, places, and times. Based on such a mixture of data, the system is able to support or critique suspicious observations in a spatiotemporal context, suggest possible alternative identifications, provide identification support based on bird characteristics, provide species-wise links to other web services (e.g., to identification documents and field guides, to bird song registries[4], and to online species identification systems), and in this way to teach the end-user in order to shorten her learning curve and to motivate her learning more. In addition, the system helps the observer in filling in data records for a legacy observation service, based on its knowledge about the context of the observation. BirdWatch is available online[5] as web application for mobile and desktop users. Additional plug-ins or application software are not needed.

In the following, the datasets, metadata model, and ontologies underlying the service are first explained. After this, an example use case of using the system is presented illustrating the functionalities of the system, and our prototype implementation is discussed shortly. The system is in trial use on the web. In conclusion, the contributions of the paper are summarized, related work is pointed out, evaluation strategies for the system are discussed, and directions for further research are outlined.

2 Data, Metadata, and Ontologies

This section explains the data, metadata, and ontologies used in BirdWatch.

[3] http://www.w3.org/standards/semanticweb/
[4] See, e.g., http://xeno-canto.org/
[5] See http://demo.seco.tkk.fi/birdwatch/. The service contains observation data only within Finland.

2.1 Observational Data

The data underlying the prototype comes from the GBIF Data Portal[6], hosting over 396,000,000 observations gathered all over the world. Our focus is on the Tiira dataset of Birds, based on the Finnish Tiira service[7] created by BirdLife and some 30 national birdwatching associations in Finland. The Tiira dataset contains 7,800,000 records. For demonstrational purposes, we selected recent data during 2007–2011 (5 years) and picked up 250,000 observations per year randomly, totaling in 1.25 million data records.

Table 1. Metadata Element Set for Observations

Element	Meaning	Identifier	Card.	Range type	Value
Species	Observed species	hh:scientific_name	1	taxmeon:TaxonInChecklist	URI
Place	WGS84 latitude	geo:lat	1	Literal	string
	WGS84 longitude	geo:long	1	Literal	string
Date	observation date	hh:date_collected	1	xsd:date	string
	Day of the year	owl-time:dayOfYear	1	xsd:nonNegativeInteger	1–366
Additions	Species in NatureGate	hh:general	0..1	Boolean	true/false
	Misindetifications	envirofi:hasCommonMisidentification	0..n	taxmeon:taxonInChecklist	URI

2.2 Metadata Model

The metadata was available in CSV format and was transformed into RDF in order to create a "5-star" linked data publication of it [3]. As a platform, the SAHA-HAKO system [5] was used and developed further[8] (e.g., the system is now directly based on a SPARQL endpoint for modularity). SAHA-HAKO creates automatically an editing environment for data with data validation functions, a faceted search engine based on the data, and a SPARQL endpoint for utilizing the data in a flexible way in applications. The RDF-based metadata model used in BirdWatch is shown in Table 1, using the namespaces below:

```
geo:      <http://www.w3.org/2003/01/geo/wgs84_pos#>
hh:       <http://www.hatikka.fi/havainnot/>
taxmeon:  <http://www.yso.fi/onto/taxmeon/>
owl-time: <http://www.w3.org/TR/owl-time/>
envirofi: <http://www.yso.fi/onto/envirofi/>
```

Here `geo` refers to the W3C Geospatial Vocabulary[9], `hh` to the observation database Hatikka of the Finnish Museum of Natural History[10] (FMNH), `taxmeon` to the taxonomic metaontology model of [12], `owl-time` to the Time Ontology in OWL[11], and `envirofi` to the EU FP7 project ENVIROFI[12]. Each observation

[6] http://data.gbif.org/welcome.htm
[7] http://www.tiira.fi/
[8] The source code is available at http://code.google.com/p/saha/
[9] http://www.w3.org/2005/Incubator/geo/XGR-geo-20071023/
[10] http://www.luomus.fi/english/
[11] http://www.w3.org/TR/owl-time/
[12] http://www.envirofi.eu/

in the GBIF data is associated to a geolocation square of 10 km x 10 km; the data publisher has not been willing to disclose the exact coordinates of observations in order to, e.g., protect endangered species. Unfortunately, more accurate geodata was not available in GBIF for common species either.

2.3 Ontologies

The basis of the system is the Birds of the World Ontology AVIO [13] we have developed. This ontology is based on the spreadsheet data available from BirdLife and FMNH, listing all birds of the world comprehensively, including scientific, English, and newest recommended Finnish names[13] [14]. The taxonomy was completed by adding higher level taxa (27 orders, 1 class and 1 kingdom) into the system obtained from the taxonomic database[14] of FMNH. This data was transformed into an ontology based on the TaxMeOn metaontology model [12] and is available as open data and as a public service in the ONKI Ontology Service[15] [11, 15]. The final AVIO ontology contains 9,740 species, 1,227 genera, and 194 families, defining the class of birds. Also a SKOS version of the ontology was created, where AVIO was extended with a corresponding vernacular namelist for the Swedish names of birds[16].

When porting AVIO to BirdWatch, some modifications and extensions to the AVIO ontology were made:

1. Tiira data uses in some cases older names for some species. These were added into the ontology manually as alternative names.
2. The ontology was enriched with envirofi:hasCommonMisidentification properties identifying similar looking species that are easily mixed. This work was based on an authoritative field guide [9] and was done by an experienced amateur ornithologist.
3. A mapping to bird species presented in more detail in the NatureGate service[17] was created. This facilitates linking BirdWatch and NatureGate services species-wise.
4. An extension to AVIO specifying characteristics of bird species was created, based on the characteristics system used in NatureGate. This system classifies birds, in terms of four major facet categories: 1) Date and location (nesting habitat), 2) Coloring and markings, 3) Shape and size, and 4) Behavior. These categories are further classified into hierarchies of subcategories. For instance, Shape and size contains subcategories for Size, Wings, Legs, Beak, Chest, Neck, and Tail on the next level. Finally, each species can be characterized by a set of values taken from the most specific categories. The identification of species can be performed as faceted search (cf., e.g., [10,

[13] http://www.birdlife.fi/lintuharrastus/nimisto/Maailman-lintujen-suomenkieliset-nimet–systemaattinen-osa.txt
[14] http://taxon.luomus.fi/
[15] http://onki.fi/en/browser/overview/linnut
[16] http://www.luomus.fi/julkaisut/muut/lintunimet/lintunimet-ruotsinkieliset.txt
[17] http://www.naturegate.fi/

2]) using the four major classification schemes as facets. There are currently 141 categories in the four hierarchical facets, such as "Short and sharp beak" and "Main color brown". In our prototype, the facets used in NatureGate were used as they are for interoperability.
5. Since BirdWatch is used for observations in Finland, AVIO ontology was pruned for this application case by removing, e.g., tropical and Australian species from it.

The bird characteristics system can be used for not only search (as in NatureGate), but also for identifying automatically potential misinterpretations between species, and point out how the species are different. For example, the characteristics of the arctic tern and common tern are quite similar with small differences regarding, e.g., beak coloring (common tern typically has some black there). A challenge in using characteristics of birds for identification is that they depend on the age of the individual, visual lighting conditions, season, and other changing factors. However, pointing out possible characteristics that may identify and differentiate bird species is the method used in guide books and is the basis for learning to identify species.

In the current version of our prototype, the bird characteristics extension to AVIO has not yet been used for automatic misconception identification. Instead, the common misconceptions links added into the AVIO ontology are used. An interesting further research question is how well misconceptions could be derived automatically based on the faceted characteristics system by, e.g., supervised learning, and whether the system after this could be used for identifying additional useful misconceptions.

3 Use Case Example

This section illustrates BirdWatch functionalities by a use case.

Assume that Olly Observer sees a bird that looks like an arctic tern near Helsinki on May 1. He would like to report about this to the Tiira system because in his mind this could be a rare observation worth reporting at the given time.

Olly opens BirdWatch page on the web with his mobile phone, and the system asks permission for positioning. He accepts this, and system pre-fills the observation form with coordinates and the current date (cf. Fig. 1). Olly then starts writing in data for the *Species* field "a..r..c..", and the system quickly autocompletes this into the full name "arctic tern". After this there are two options to proceed: 1) Pushing the *Check* button would retrieve supporting and critiquing information for the hypothetical observation. 2) Send button would send the data to the Tiira.fi service without providing such information—in this case Olly should be confident about the identification. Olly decides to push *Check* because he is not quite sure about the bird species, and the system provides him with the following information for consideration under the *Check* button, cf. Fig. 1:

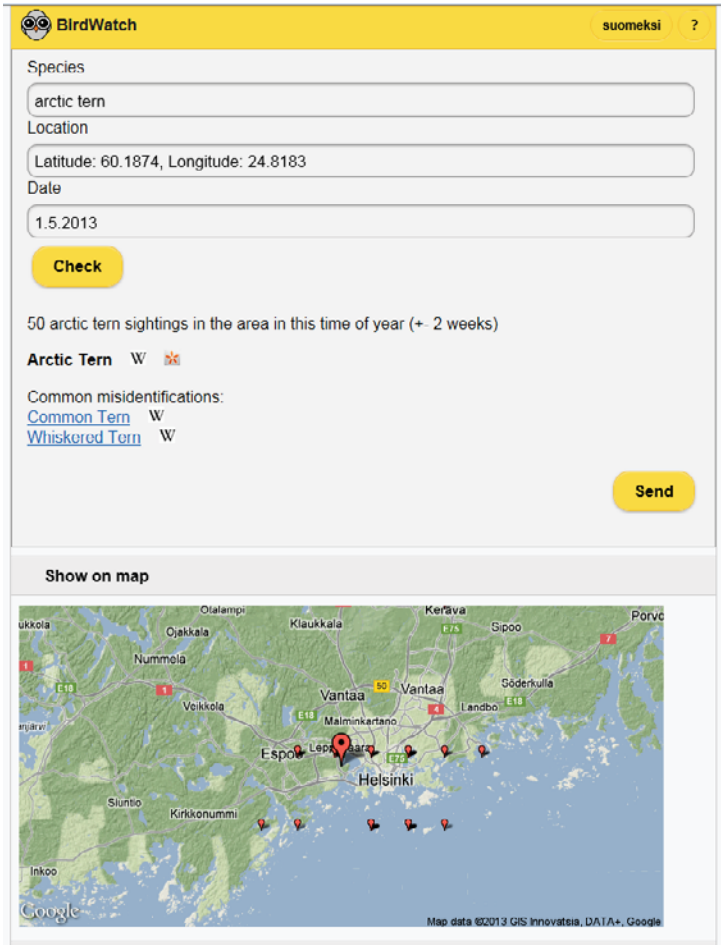

Fig. 1. Screenshot with a map showing the current position and related observations.

1. The number (50) of similar observations in the given area and within a time interval of two weeks before or ahead is presented. A low number can be considered a warning of possible misinterpretation. In this case, however, the number 50 suggests that the observation is not particularly rare and definitely possible.
2. Links to recommended identification services on the web for the arctic tern are provided, here links to the arctic tern pages of Wikipedia and NatureGate (indicated by special button symbols).
3. Links to commonly misinterpreted species of the arctic tern are provided, in this case the common tern and the whiskered tern. By following these links, Olly can change the proposed observation and get new statistics for

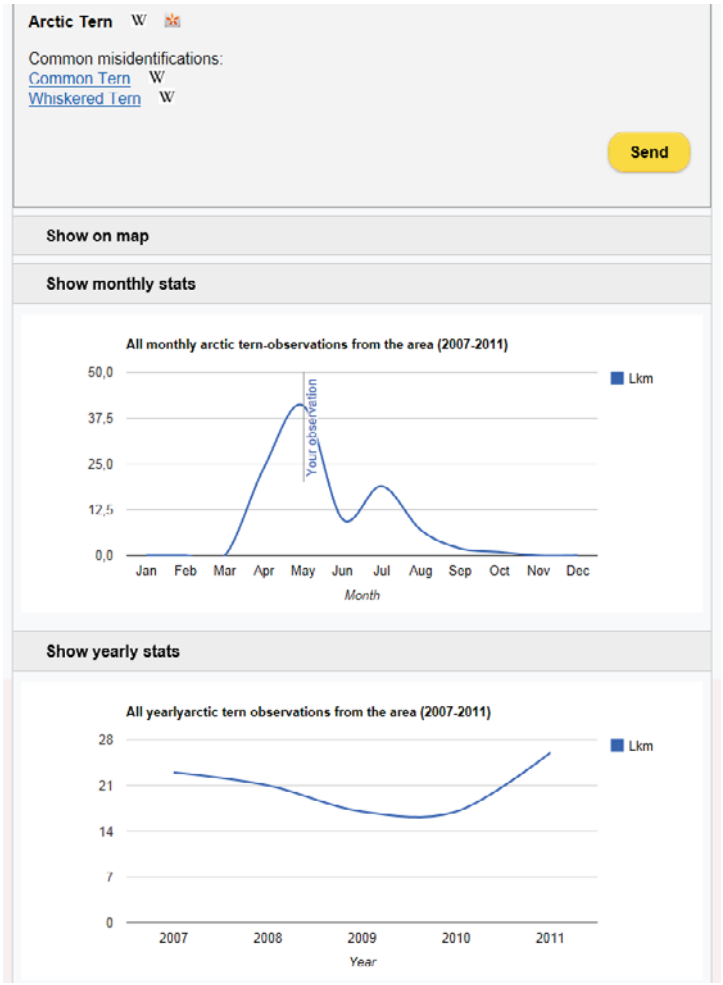

Fig. 2. Screenshot with statistical info related to the observations.

these species and other information. In this case, the statistics would tell that there are no observations made in this place and time about whiskered terns, so this option would be unlikely (although possible). However, there are 192 observations of the common tern, raising the question whether Olly actually saw a common tern.

4. After each listed possible misconception there are links to further information services on the web. In our prototype, links to species pages in Wikipedia and NatureGate are provided, but the list can be extended to other sources, too, such as online bird song registries, based on linked data. Olly can find

out in this way characteristics differentiating the common and arctic terns, if he does not recall them otherwise.

5. Arctic tern observations on the map are shown, centered around the current location. The idea here is that observations nearby may support the current identification hypothesis or be against it.
6. A monthly statistics of arctic tern observations at the spot in the given time frame is shown (cf. Fig. 2). This visualization complements information about possible misconceptions in time: it may be the case that even if similar observations were rare at this point in time, the situation could change radically due to, e.g., migration soon. In this case, the situation is indeed very dynamic—the number of observations increases quickly during April—but on May 1 the peak has already been reached.
7. Also a yearly statistics of all arctic tern observations in the area is shown. This is an interesting piece on knowledge from a biodiversity point of view and could be of interest to Olly. A raising and high number of similar observations indicates that the species is generally not rare and provides some support that the hypothesized observation is feasible.

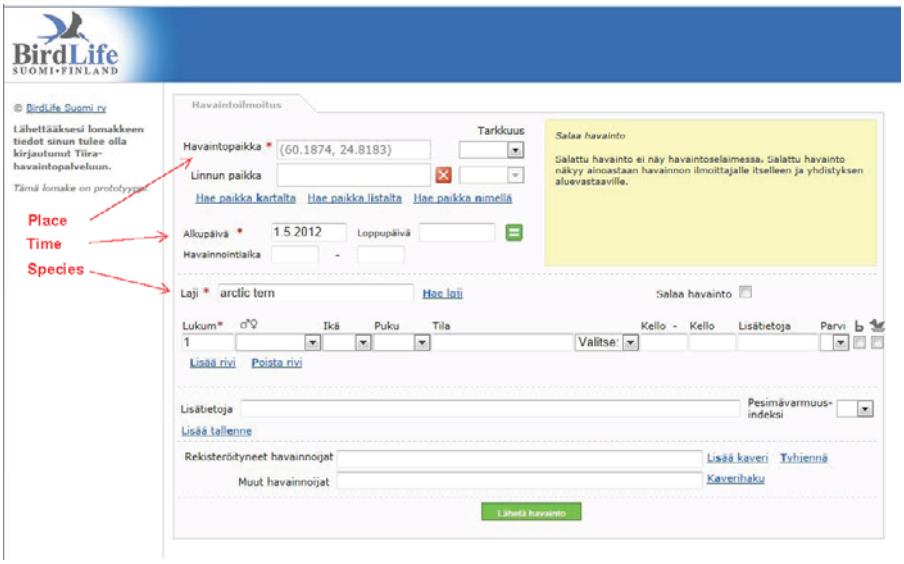

Fig. 3. Pre-filled observation form for the Tiira service.

Given the information listed above, Olly thinks that the observation hypothesis arctic tern seems to be correct and pushes the *Send* button to submit the observation into the Tiira system. The system then fills in partly the legacy Tiira

observation submission form[18] as depicted in Fig. 3, including, e.g., fields for the species, time, and location. In this way the time needed for filling up the data is shortened, which saves time for the actual observation work. Obviously, a less time consuming reporting system also motivates end-users to actually submit their observations. In the prototype, we use the Tiira legacy reporting form as it is, designed for desktop devices. Designing a better interface for mobile devices remains a topic for further development.

In short, the system provides statistical knowledge in context for evaluating the feasibility of the proposed observation, ontology-based information about possible misinterpretations, links to additional web services that may help in the identification and for learning more about the species, and finally speeds up reporting by pre-filling observational reports.

Some fields of the data record can be filled automatically based on the context of observation (e.g., place and time), for others, such as species reference, ontology services [11, 7] can be used for finding and fetching the right URIs.

4 Implementation and Visualization

The BirdWatch prototype is an HTML5 Mobile application that is implemented using JQuery Mobile[19]. The autosuggestion of species, recommendation links, and visualizations are created and queried directly from the underlying linked data SPARQL endpoint using Ajax requests. The application uses W3C Geolocation API[20] for detecting the location of the user. The user can also position the observation by inputting the name of the location or address that is then processed with the Google Maps API[21]. Once the user has given the input about the species, and the location of the observation and time are known (also time can be changed manually), a SPARQL query is send to the observation triplestore and the observations of the given area and time are analyzed and visualized.

The fuzzy locations (+-10km) of the observation data are plotted on a map, and details about the observations in the area are processed from the JSON serialization of the SPARQL response using the same method as in sgvizler [8]. The query results are transformed into a format used by the Google chart library[22], and represented as a graph visualizing the fluctuations of observation data on a monthly and yearly basis.

5 Discussion and Evaluation

Species distribution maps for different times (e.g., for nesting time and overwintering) are widely used for species identification, and maps are available in field

[18] http://www.tiira.fi/, the web form is available in Finnish and Swedish
[19] http://jquerymobile.com/
[20] http://www.w3.org/TR/geolocation-API/
[21] https://developers.google.com/maps/
[22] https://developers.google.com/chart/

books, such as [9]. Online systems, such as eBird[23] used by, e.g., the Audubon Society, provide online visualizations of observations, such as range and point maps and yearly bar charts. Different metrics of observations can be graphed along a timeline and statistics of one species contrasted with others. Data mining tools can be applied to observational databases in order to analyze and discover phenomena that take place in the nature [4]. There are charactetistics-based mobile bird identification systems online, such as WildLab-Bird[24], iNaturalist[25], Project Noah[26], and NatureGate[27], aiming at teaching birdwatching to citizens and at the same time collecting observations.

BirdWatch makes use of GBIF data and its metadata model[28] (based on Ecological Metadata Language EML) that is transformed directly into RDF. Other metadata formats and vocabularies used for describing observational data include, e.g., Darwin Core[29] and OBOE OWL[30]. In our case, there was no need for complex modeling since the underlying data available was simple GBIF data. As for species ontologies, related work includes the TaxonConcept project[31], focusing on aggregating and linking taxon data from different sources. Numerous scientific name repositories[32] are in use in biology and can be used as a basis for species ontologies—we used the name list of BirdLife and the translations of common names from FMNH since they focus on birds.

The novelty of BirdWatch regarding these systems is based on the following ideas: The visualizations are provided in the *spatio-temporal observation context*, based on an proposed observation. Our goal is to help the observer to improve data quality rather than just provide visualization or data mining tools for inspecting the data for, e.g., research purposes. Furthermore, BirdWatch is arguably the first birding support system to use ontologies and the Linked Data approach: our approach therefore has the potential of not only use statistics but also structured knowledge to explain characteristics of birds, identify common misinterpretations between species, and link observation candidates to additional online services, such as identification assistants, Wikipedias, sound registries, and other observation services. The Linked Data approach has been proven useful when aggregating data from distributed, heterogeneous observation repositories in an interoperable way in many fields of application.

A system such as BirdWatch needs to be evaluated at least along the following dimensions: 1) computational efficiency, 2) ease of use, and 3) capability of raising data quality. As for computational efficiency (1), our experiment suggests that using a SPARQL endpoint as a basis scales well up to at least millions of

[23] http://ebird.org/
[24] http://bird.thewildlab.org/
[25] http://www.inaturalist.org/
[26] http://www.projectnoah.org/
[27] http://www.naturegate.fi/
[28] http://www.gbif.org/informatics/discoverymetadata/ipt-and-metadata/
[29] http://rs.tdwg.org/dwc/
[30] https://semtools.ecoinformatics.org/oboe
[31] http://www.taxonconcept.org/
[32] http://gni.globalnames.org/data_sources

observations using ordinary triplestore tools and hardware. The system could also be implemented using, e.g., a REST API (JSON) on a standard database system that scales up even better. However, relational databases are not as flexible as SPARQL triplestores for data aggregation, linking, and querying. Ease of using the interface (2) of the prototype has not been tested systematically, but a few test ornithologists have tried the system out without major difficulties. The interface is in any case quite simple, and pre-filling the Tiira observation form of course helps in reporting without an additional burden. However, we envision that understanding and interpreting the statistical data may be an issue when using the system. One test user, for example, asked why she did not find an observation that she had made earlier in Tiira. The problem was that the data in the system is not updated in real time, but harvested with latency from GBIF. The system should of course be integrated with the actual Tiira system in real time, but this has not been done yet in the demonstration system.

Another issue is that the underlying observational data is by no means complete and it is biased in many ways, because it is based on the observations of the public. For example, consider the monthly statistics of swallows. In springtime there are lots of early reports of the first swallows seen in Finland, but in summertime people lose interest in them because they are quite common and are seen virtually everywhere in southern Finland. The statistical monthly curve therefore goes down but this does not really tell us how common swallows are in summer but only about the number of reported observations. The user must understand this and interpret the data correctly, otherwise the data may guide her to false or too conservative interpretations. The situation is different when using professional surveying datasets where all birds seen are systematically and reliably reported during a time period and within an area. Our approach and system could of course be applied to such datasets, too, by adjusting the interpretation of statistics.

The most difficult evaluation task is to measure whether using a system like BirdWatch actually improves data quality (3) in the long run and how. One possibility to measure this would be to select a set of test users, and record and evaluate their experiences in using the system. For example, the test users could mark up situations where they think the additional information was helpful in some way, e.g., in preventing making an interpretation that after a second thought was wrong. Even if the final objective truth of the observation could not be verified for sure, subjective measurements of this kind would be helpful in determining the usefulness of the system in raising the quality of observations. Evaluating the system in such a setting remains a topic for further research.

Acknowledgements This work is part of the National Semantic Web Ontology project in Finland[33] FinnONTO (2003–2012), funded by the National Technology and Innovation Agency (Tekes) and a consortium of 35 public organizations and companies. Support from the Linked Data Finland project[34] and the EU

[33] http://www.seco.tkk.fi/projects/finnonto/
[34] http://www.seco.tkk.fi/projects/ldf/

FP7 project ENVIROFI[35] is acknowledged, too. Thanks to BirdLife, Finnish Museum of Natural History, and NatureGate for fruitful collaboration.

References

1. Gaston, K.J., Spicer, J.I.: Biodiversity: an introduction. 2nd Ed. Blackwell Publishing, Oxford, U.K. (2004)
2. Hearst, M.: Search User Interfaces. Cambridge University Press, New York (2009)
3. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136, Morgan & Claypool (2011)
4. Hochachka, W., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., Kelling, S.: Data-mining discovery of pattern and process in ecological systems. Wildlife management 71(7), 2427–2437 (September 2007)
5. Kurki, J., Hyvönen, E.: Collaborative metadata editor integrated with ontology services and faceted portals. In: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010. CEUR Workshop Proceedings, Vol. 596 (June 2010)
6. Levin, S.A. (ed.): Encyclopedia of Biodiversity. 2nd Ed. Elsevier Science Publishing (2013)
7. Noy, N.F., d'Aquin, M.: Where to publish and find ontologies? A survey of ontology libraries. Web Semantics: Science, Services and Agents on the World Wide Web 11(0) (2011)
8. Skjæveland, M.: Sgvizler: A JavaScript Wrapper for Easy Visualization of SPARQL Result Sets. In: Extended Semantic Web Conference 2012 (ESWC2012). Heraklion, Crete, Greece (May 2012)
9. Svensson, L., Mullarney, K., Zetterstöm, D.: Fågelguiden—Europas och medelhavomrodets fåglar i färg. Bonnier, Stockholm, Sweden (1999), English edition: Collins Bird Guide, HarperCollins, London, 2000
10. Tunkelang, D.: Faceted Search. Morgan & Claypool, Palo Alto, USA (2009)
11. Tuominen, J., Frosterus, M., Viljanen, K., Hyvönen, E.: ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In: Proceedings of the 6th European Semantic Web Conference (ESWC 2009). Springer–Verlag, Berlin (2009)
12. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological names and taxonomies on the semantic web – managing the change in scientific conception. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011). Springer–Verlag, Berlin (2011)
13. Tuominen, J., Laurenne, N., Koho, M., Hyvönen, E.: The birds of the world ontology avio. The 10th Extended Semantic Web Conference (ESWC 2013), Proceedings of the Poster Papers (2013)
14. Väisänen, R.A., Högmander, H., Björklund, H., Hänninen, L., Lammin-Soila, M., Lokki, J., Rauste, V.: Maailman lintujen suomenkieliset nimet (Finnish names of the birds of the world). 2., uudistettu painos (2nd edition)
15. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The Finnish ontology library service ONKI. In: Proceedings of the 6th European Semantic Web Conference (ESWC 2009). pp. 781–795. Springer–Verlag, Berlin (2009)

[35] http://www.envirofi.eu/

iPlant SSWAP (Simple Semantic Web Architecture and Protocol) enables semantic pipelines for biodiversity

Damian D. G. Gessler[1], Blazej Bulka[2] , Evren Sirin[2], Hans Vasquez-Gross[3], John Yu[3], Jill Wegrzyn[3]

[1]The iPlant Collaborative, University of Arizona, Tucson, AZ, U.S.A.
dgessler@iplantcollaborative.org
[2]Clark and Parsia, Washington, D.C., U.S.A.
{blazej,evren}@clarkparsia.com
[3]University of California, Davis, CA, U.S.A.
{havasquezgross,jjsyu,jlwegrzyn}@ucdavis.edu

Abstract. Real-time response is a basic characteristic of the Web. Yet semantic reasoning at transaction-time supporting real-time response remains challenging. Here we report how the iPlant Semantic Web Platform uses SSWAP (Simple Semantic Web Architecture and Protocol; http://sswap.info) for transaction-time reasoning, service discovery, workflow construction, and execution. The platform enables users at web sites, such as TreeGenes' DiversiTree and CartograTree, to select data and use it for real-time semantic discovery into a knowledge base of semantic web services. The platform uses first-order, description logic reasoning and just-in-time ontologies to allow users to drag-*n*-drop independent, distributed semantic web services into a semantic pipeline. This enables biodiversity research using data sets from TreeGenes, FLUXNET (Ameriflux), WorldClim, and TRY-DB integrated under a common web front-end called CartograTree. Scientific use cases are for tree scientists to associate phenotype and/or environmental traits with underlying genotypes in geo-referenced forest trees across a distribution of Web resources.

Keywords: Semantic Web Services, SSWAP, iPlant, TreeGenes, DiversiTree, CartograTree, OWL

1 Introduction

Bioinformatic software exhibits long-tail characteristics: a relatively small number of programs and web sites are widely used (*e.g.*, [1,2]), while a much

larger number are used by varied audiences for specialized applications (*e.g.*, [3,4]). The iPlant Collaborative [5] seeks to enable data-driven scientific integration, both within the enterprise and across Web resources, including widely used programs of general interest and niche programs for specific needs. Emphasis is on having software layers handle data and service syntax and semantics (including independently developed and maintained long-tail offerings) thereby freeing the scientist to focus on data and service discretionary use. To achieve this, iPlant is using SSWAP (Simple Semantic Web Architecture and Protocol [6]) in a drag-*n*-drop semantic pipeline motif with third-party Web site integration. In this paper we report how a collaboration with TreeGenes [7] enables biodiversity applications in forest genetics. Exemplary applications in land management and biodiversity include the identification of specific genotypes that may be best suited for reforestation, or the development of strategies for tree migration as it relates to climate change. In both cases, genotypes that influence traits such as cold-hardiness, drought-tolerance, and disease resistance can be examined in relation to environmental characteristics of target regions including elevation, soil composition, and precipitation.

2 The Platform

Architecture The iPlant Semantic Web Platform is a Web architecture of four distributed actors: *i*) providers of services; *ii*) consumers of services; *iii*) ontology severs; and *iv*) a semantic Discovery Server (pipeline-maker and match-maker). Data—be it unstructured, semi-structured, or structured (*e.g.*, as in relational database stores)—enters the system via a service interface layer; *i.e.*, the platform does not operate on raw data *per se*, but via service interfaces, the invocations of which yield access to, and transformations of, data. This service interface layer is key to enabling distributed data to be integrated "rationally" under a first-order description logic protocol.

SSWAP (Simple Semantic Web Architecture and Protocol) SSWAP is a 100% W3C OWL DL-compliant light-weight protocol of five classes and 12 properties. It allows services to describe what they are, the types of data they consume, and the types of data they produce. The protocol's ontology in its entirety is at [8]. The five classes correspond to: *i*) the service Provider, *ii*) the service itself, called a Resource, *iii*) a data structure construct called a Graph, *iv*) input data (a Subject), and *v*) output data (an Object). SSWAP is the service analog of the fundamental RDF data model of mapping a subject to an object via a property; in the cases of SSWAP, the protocol maps a Subject to an Object via the implicit operation of a service (the Resource). Subject and Object instances may be URIs, thereby allowing for indirection and non-serialization of data, or

they may identify data structures of arbitrary OWL sub-graphs, with properties and serialized data. Instances of Resource, Subject, and Object may be annotated with user-defined ontologies and thus are "unlimited" in domain scope; the protocol simply defines the scaffold. Services may have multiple Subjects mapping to multiple Objects. A protocol description of a service is called an RDG (Resource Description Graph). An HTTP GET on the Resource URL of the RDG returns the RDG in W3C-compliant OWL RDF/XML. Because service descriptions are just text documents retrievable by a simple GET, they are readily available for search engine traversal and viewable by browsers[1]. An RDG with input data creates an RIG (Resource Invocation Graph). An HTTP POST of the RIG or a GET with ontology *term=value* assignments in the query string invokes the service. An RIG with output data is called an RRG (Resource Response Graph). Thus SSWAP creates an ecosystem of protocol graphs, all sharing a canonical model, with a common syntax (OWL RDF/XML), under a common services' semantic (SSWAP), amendable to customization by user semantics (adding ontology terms to the Resource, Subject, and Object). SSWAP is a wrapper technology, so it can semantically enable legacy and non-semantic services. Notably, a SSWAP service *description* yields the service amenable to automated semantic *discovery*, *invocation*, and *response*.

Semantic Querying A service's protocol description encapsulates the information needed for its discovery and invocation. Thus one can consider any putative RDG as a query graph (called an RQG: Resource Query Graph) into a knowledge base of all RDGs. For semantic querying, we find all services for which the RQG's: *i*) Resource is a subclass, and *ii*) Subject is a super-class, and *iii*) Object is a subclass, of any service in the knowledge base. Subsumption reasoning covers arbitrary complex, inferred, anonymous classes. The resultant services, and only these services, are guaranteed to be of the type of service queried (or more specialized), to operate on the input data (or generalizations of it), and return data of the requested output type (or specializations of it). This allows us to use a reasoner for match-making based on the output of one service being logically sufficient for the input of another. Thus reasoning is used to examine service descriptions, input data types, and output data types, to enable semantic matching with published services.

Constructing semantic pipelines At `http://sswap.info`, a Web front-end to a backend pipeline manager allows users to connect services into pipelines. Pipelines are built on-demand by using transaction-time reasoning to aid the user in building a workflow of distributed services.

Start with a lexical search Users at `http://sswap.info` may search for services using keywords. Upon selecting a service and adding it to a new

[1] Visit `http://sswap.info`, search for a service, click on 'Service URI' to view the RDG, or visit `http://sswap.info/api/makeRDG` for examples.

pipeline via web-based drag-*n*-drop, we present the user with all downstream services that can operate on the upstream service via semantic querying as described above. In this manner, the user can build a pipeline of services. For each service, we reason over the service's RDG to determine its necessary and sufficient conditions, and based on this construct on-demand a custom user dialog that allows the user to enter the service's required and optional parameters, if any. In a similar manner the Subject is examined, and the user may upload data to be ontologically tagged via the RDG. The protocol declares a datatype property *sswap:inputURI*[2] which allows service providers to write custom Web pages to solicit user input for their services. If *sswap:inputURI* resolves to a Web page, the platform will present that page to the user in addition to allowing the user to use the auto-generated, custom user dialog.

Start with data launched from a web site We provide a Javascript snippet that allows any webmaster to add a "sswap.info" button to their web pages. We call this Web Discovery. We provide a service to allow the Web master to package or reference the data using JSON (see `/api`)[3]. Upon the user pressing the Web Discovery button, the JSON is sent to our Discovery Server, where we translate it into an RDF/XML RQG, perform semantic querying, and present the user with a new pipeline preloaded with their data and the semantic results of all matching candidate downstream services.

Start with the results from previous pipelines Because the last service in a pipeline returns a standard RRG, this can be used to start a new pipeline. In this manner, a pipeline can seed new pipelines. Data is private, but pipelines may be published for public use and are semantically discoverable like services. In this manner, we grow a database of user-built combinations of Web distributed services; this has deep social networking value. We note that public sharing of pipelines does not imply unregulated execution of services: any service is free to gate-keep resources with logins, HTTPS, and so forth.

Pipeline invocation is orchestrated, but execution is distributed RDGs represent published SSWAP services that are offered by third-parties on the Web. When the user initiates a pipeline, we coordinate the invocation and callback of services, but do not ourselves execute the services: the services run independently, asynchronously on their host machines. Downstream services retrieve the upstream RRG from the upstream service with a token and convert it to an RIG without passing through our servers. In this manner we are not privy to non-serialized data being transferred between services, thereby maintaining an important privacy safe-guard.

Transaction-time reasoning SSWAP graphs (RDGs, RIGs, RRGs, and RQGs) are small documents of a few dozen lines of W3C OWL [DL]

[2] *sswap:* prefix resolves to `http://sswapmeet.sswap.info/`
[3] Relative URLs are RESTful endpoints on `http://sswap.info/`

RDF/XML that typically expand to a few thousand triples after first-order reasoning. We use reasoning in four places: *i*) when Providers publish their RDGs with us, we resolve ontology terms by dereferencing them on the Web; we then infer over the closure RDG and store the resulting inferred graph in a triple-store [9]. We use a combination of transaction-time reasoning at publication time and offline processing to maintain the knowledge base; *ii*) when users initiate Web Discovery from a web site by sending us a JSON representation of an RRG, we resolve the RRG, convert it to a RQG, and execute transaction-time semantic querying; *iii*) when users build pipelines we reason during the transaction process to satisfy semantic querying and other pipeline duties; *iv*) when third-party services receive an RIG they need to process the request and return a RRG that complies with the logical contract of their RDG. We provide a kit (`/sdk`) that allows third-parties to run their own servlet reasoner to handle transaction-time reasoning to process requests.

Pipeline management Control is architected as three separate components: *i*) we use Vaadin [10] to offer a RIA (Rich Internet Application) enabling an intuitive, drag-*n*-drop user experience; *ii*) communication to the backend is performed by a 100% RESTful JSON API, making heavy use of idempotent HTTP GETs and PUTs. This means that a user may start building a pipeline, bookmark it, close their browser, and open it anywhere, anytime, and continue their work. It means that users may begin long-running pipelines, and return at their convenience with a different browser and Web session; *iii*) the pipeline manager communicates with the Discovery Server via a RESTful API.

Platform APIs We wrote ~185,000 lines of open-source Java code to build a platform, Java API (`/javadocs`), and helper services. We use the Java API internally, and package it as part of our SDK (Software Development Kit) so anyone may write their own SSWAP services (`/sdk`). Many developers are fluent in JSON, but not in OWL RDF/XML, so we wrote a RESTful translator that allows SSWAP graphs and user ontologies to be written in JSON and then translated to OWL RDF/XML (`/api`; see also `/make` and [11]). We expose Discovery Server engagements as RESTful endpoints (`/wiki/api`).

3 Ontologies

A challenge for the semantic web services is how to enable and incorporate distributed ontologies. We enable the use of user-defined OWL ontologies to allow services to describe their data, and to allow clients to query and engage said services.

Just-In-Time ontologies We used Smart GWT [12] to write an application that allows anyone to host their ontologies on our servers [11]. Users register for a free iPlant account and may create and administer new ontologies (called

"namespaces"). Users build ontologies term-by-term using a JSON syntax [13], translate them to RDF/XML with the press of a button, and publish them on-demand. Terms are separately dereferenceable and immediately available to anyone on the web. Just-In-Time ontologies lower the barrier to entry for creating and using small, agile ontologies, but they are not required: ontologies residing anywhere on the web may be freely used, subject to byte and time limits during transaction processing. Ontological statements (*e.g.*, definitions and relation to other terms) are read and used in reasoning if dereferencing term URIs returns OWL RDF/XML statements.

Support for "large" legacy ontologies: module extraction with BioPortal BioPortal [14] is a major repository funded by the National Center for Biomedical Ontology. It contains over 320 ontologies, and over 180 OWL ontologies. We use the method of [15,16] to process each OWL ontology offline to generate "atoms," such that at transaction-time we can compute the subset of statements (called a "module") that are necessary and sufficient for complete entailment over any subset of terms (called a "signature"). Importantly, for moderate sized signatures the module is often much smaller than the ontology itself [15], thus lending it as a key approach to bringing large, legacy ontologies to transaction-time applications in the semantic Web. Currently, ontology modularization is available as a service at `/modularize`. As of this writing, we are implementing a strategy to incorporate it into transaction-time processing but this is not yet part of the larger platform.

Ontologies enable semantic querying and reasoner-assisted semantic pipeline construction When we process a SSWAP graph, we extract ontology terms and dereference them to retrieve their OWL statements. If these documents themselves contain terms, we dereference those, and continue this cascade until closure is achieved, subject to traversal depth, byte, and time limits. For Web Discovery and pipeline construction we then use Semantic Querying (described above) to find matches between data and/or the output semantics of the upstream service and the input semantics of all putative downstream services. Subsumption determination is performed at transaction time, so axiomatic subsumption claims (*e.g.*, `rdfs:subClassOf`) are supported but not required: the reasoner uses transaction-time classification to determine subsumption. Note that it is the SSWAP protocol that makes this possible, because the protocol ensures that the subject and object semantics of RDGs, RIGs, RRGs, and RQGs are comparable.

7

4 Integrating Enterprise, HPC, and the Semantic Web for Biodiversity

Enterprise resources TreeGenes [7] is a large biological resource serving over 2500 forest geneticists from over 800 organizations. It contains data from 15 yrs on over 1200 species, including genomic, phenotypic, and other data. We wrote 11 SSWAP services to expose slices of this data and added SSWAP Web Discovery to TreeGenes' DiversiTree [17]. For geographically-oriented tree scientists, we wrote a mapping tool called CartograTree [18,19]. Researchers can search specific geographic regions, tree species, phenotypes, or environmental parameters and customize their analysis accordingly. We enabled CartograTree with SSWAP Web Discovery so that scientists can launch directly into semantic discovery. The iPlant Collaborative serves over 7500 scientists with enterprise-class and High Performance Computing (HPC) resources, petabyte-scale storage, and other resources. We wrote semantic pipeline support to engage HPC XSEDE resources [20] and used SSWAP to semantically wrap 10 resources in the domain of multiple sequence alignment and phylogenetic tree reconstruction.

Biodiversity The DiversiTree/CartograTree/SSWAP integration is driven by questions arising from climate change, disease resistance, and conservation. Knowledge of the adaptive genetic potential of forest tree populations is critically important for evaluating their vulnerability to a changing climate [21]. Forests are key to sequestering carbon and consequently contribute an important role to mitigating or reinforcing the impacts of climate change. Healthy forests provide fundamental habitat for valued biodiversity and essential ecosystem services in the form of global carbon cycling, clean water and air, fiber, and recreation. Sustaining healthy forests in the face of climate change is a central challenge for resource management [22]. Towards this goal, researchers are examining candidate loci to understand how individuals and populations are impacted by environmental factors. Specifically, a fusion of population genetics and landscape ecology to layered geographic information systems allows for focused studies of how landscape features affect genetic variation [23-25].

Experimental design often focuses on first identifying candidate genes under selection from geoclimatic factors, determining their allelic diversity, and testing for associations between trees' genotype, phenotype, and the environment. CartograTree connects the TreeGenes' repository of genotype and sequence data to environmental and phenotypic resources. TreeGenes houses approximately 901,000 sequences, 24 million genotypes, and 20,000 phenotypes on individuals from over 1,200 different forest tree species. Sequencing includes either Sanger-based or next-generation approaches, and used to identify polymorphisms in small populations. The polymorphisms are then validated in larger populations

through the use of high-throughput genotyping assays. In many cases, genotyped trees are phenotyped for various traits. Barcode identifiers assigned during sample collection are maintained through DNA extraction, sequencing, genotyping, and phenotyping, while also associating trees with their geo-referenced coordinates. The external sources supplying environmental and phenotypic data include relevant portions of the FLUXNET (Ameriflux) [26], WorldClim [27], and TRY-DB [28] repositories. Ameriflux represents 81 remote sensing sites across North and South America; WorldClim is a compilation of five different climate databases covering the globe; TRY-DB enhances phenotypic data with approximately 80,366 geo-referenced phenotypic records represented by 368 species. Within CartograTree, specific queries and filters are available to select by genus, species, or phenotype of interest. The phenotypic selections include economically relevant traits, disease evaluations, and developmental metrics. The map portion of the interface gives users the option to select regions of interest, and capture the associated environmental data, such as slope, elevation, precipitation, seasonal temperatures, and more. From this, scientists can send selected data for SSWAP Web Discovery, for example, to perform multiple sequence alignment and phylogenetic tree reconstruction on high performance computing clusters. A full description of CartograTree and SSWAP is published at [19]. Association studies are facilitated through the ability to create flat files based on the common phenotypic or environmental evaluations for a selection of trees. The results of these studies are aimed at improving land-management decisions through the identification of genotypes that will thrive in specific environments; information that is necessary for reforestation, disease resistance, and climate change.

5 Conclusion

Semantics and biodiversity is still in its nascent years. Our work is focused on a division of scientific labor between domain-specific information resources such as TreeGenes, infrastructural resources such as iPlant, high performance computing assets such as underlying the phylogenetic applications available on XSEDE, and the larger Web. iPlant's Semantic Web Platform is developed as the technological conduit for integration across these resources. It uses transaction-time first-order description logic reasoning to allow semantic web services to be discovered, connected, and invoked via a simple drag-*n*-drop web interface. TreeGenes, DiversiTree, and CartograTree offer an initial foray into the use of these technologies for forest tree biodiversity.

Acknowledgements We thank Pavel Klinov for work on ontology modularization and Yan Kang for work on the Just-In-Time Ontology editor.

9

This work was supported by NSF grants for the iPlant Collaborative (#DBI-0735191) and SSWAP (#0943879).

6 References

1. http://www.ncbi.nlm.nih.gov/sites/gquery
2. Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. Basic local alignment search tool. J. Mol. Biol. 215: 403-410 (1990).
3. https://pods.iplantcollaborative.org/wiki/display/DEapps/List+of+Applications
4. Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. Aug 25;11(8):R86 (2010).
5. Goff, S. A. *et al.*, The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Frontiers in Plant Science 2, 1 - 16 (2011). DOI: 10.3389/fpls.2011.00034.
6. Gessler, D.D.G., Schiltz, G.S., May, G.D., Avraham, S., Town, C.D., Grant, D., Nelson, R.T.: SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services. BMC Bioinformatics, 10:309, pp. 1-21 (2009).
7. http://dendrome.ucdavis.edu
8. http://sswapmeet.sswap.info/sswap; see also http://sswapmeet.sswap.info/jit/sswap. Also included at the URL is an additional class and property for asynchronous service invocation.
9. Pellet + Stardog; see http://stardog.com
10. http://vaadin.com
11. http://sswapmeet.sswap.info
12. http://smartclient.com
13. http://sswap.info/api/JSONSyntax
14. http://bioportal.bioontology.org
15. Del Vescovo, C., Gessler, D.D.G., Klinov, P., Parsia, B., Sattler, U., Schneider, T., and Winget, A.: Decomposition and Modular Structure of BioPortal Ontologies, In: ISWC, LNCS, 7031: pp. 130-145 (2011)
16. Klinov, P., Del Vescovo, C., Schneider, T.: Incrementally updateable and persistent decomposition of OWL ontologies. In: Proceedings of OWL: Experiences and Directions Workshop 2012. Klinov, P., Horridge, M. (eds.) Heraklion, Crete, Greece, May 27-28, 2012. CEUR Workshop Proceedings 849 CEUR-WS.org 2012.
17. http://dendrome.ucdavis.edu/DiversiTree
18. http://dendrome.ucdavis.edu/cartogratree
19. Vasquez-Gross, H.A., Yu, J.J., Figueroa B., Gessler D.D.G., Neale D.B., Wegrzyn J.L.: CartograTree: connecting tree genomes, phenotypes and environment. Molecular Ecology Resources, (2013) doi: 10.1111/1755-0998.12067
20. https://www.xsede.org
21. Neale D.B., Kremer A.: Forest tree genomics: growing resources and applications. Nature Reviews Genetics 12, pp. 111–122 (2011)

10

22. Peterson, D.L.; Halofsky, J.E.; Johnson, M.C.: Managing and adapting to changing fire regimes in a warmer climate. In: McKenzie, D.; Miller, C.; Falk, D., (eds.) The landscape ecology of fire. New York: Springer: Chapter 10, pp. 249–267 (2011)
23. Manel, S., Joost, S., Epperson, B.K. et al.: Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. Molecular Ecology, 19, pp. 3760–377 (2010)
24. Manel, S., Schwartz, M.K., Luikart, G., Taberlet, P.: Landscape genetics: combining landscape ecology and population genetics. Trends Ecol. Evol. 18, 189-197 (2003)
25. Feder, M., Mitchell-Olds, T.: Evolutionary and ecological functional genomics. Nature Genetics Reviews, 4, pp. 649-655 (2003)
26. Baldocchi, D., Falge, E., Gu, L.H., *et al.*: FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bulletin of the Amer. Meteor. Soc., 82, pp. 2415–2434 (2001)
27. Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25, pp. 1965–1978 (2005)
28. Kattge, J., Diaz, S., Lavorel, S. et al.: TRY - a global database of plant traits. Global Change Biology, 17, pp. 2905–2935 (2011)

A knowledge base for Exploited Marine Ecosystems

Julien Barde[1], Pascal Cauquil[1], and Norbert Billet[1]

[1] Institut de Recherche pour le Dévelopement, UMR EME 212, Sète, France

Abstract. In 2008, IRD started to work on setting up a knowledge base (named Ecoscope) about Ecosystem Approach to Fisheries domain (EAF) in the context of a marine ecology laboratory studying Exploited Marine Ecosystems in different regions of the world. This application was meant to fit the needs of researchers by improving knowledge and related information resources management [14,12]. Among other goals, researchers expected an information system enabling to provide an inventory of available data sources (ecological observations, satellites images, pictures, articles, reports..) and facilitating data rescue, data access, data processes (indicators..) as well as the ability to summarize related knowledge through fact sheets about domain entities (ecosystems, species..) connected with hyperlinks (based on ecosystem relationships).
Beyond metadata, data management and related interoperability issues (OGC, TDWG...), this project was then a real opportunity to set up an ontology for EAF domain in order to link existing information resources with real-world entities (EAF domain concepts). To achieve these goals, semantic Web standards and reference RDF schemas have been taken into account (SKOS, Dublin Core, FOAF, OBOE, Darwin Core...) and a first version of RDF schema for EAF domain has been set up. These ontologies have been instantiated to describe our information resources and some knowledge about entities that researchers are studying.
A first Website has been set up on top of this knowledge Base. Related Web pages consist mainly in fact sheets about domain entities (ecosystems, top predators and related preys species, fishing vessels..persons) where users can find related information resources (spatial layers, articles, pictures, indicators..). Knowledge can as well be summarized through networks of entities like food webs with dedicated visualizations tools. This is made possible by querying the knowledge base where Linked metadata and data (in underlying databases) are tagged with related species URIs. Proof has been done that Semantic Web languages can be used to fit the needs of our colleagues. Moreover, in the context of iMarine FP7 project, we started to deal with partners having similar projects (FLOD from FAO, Worms, FORTH). We then set up a SPARQL endpoint and OpenSearch access to share the content our knowledge bases with other applications (search engines, text mining applications..).
We will present our current application and related technical choices as well as futur plans to connect additional data sources to enrich this knowledge base and make it available for our partners. In particular we will describe some use cases related to biodiversity management issues.

1 Ecosystem Approach to Fisheries

In this paper, we present our work on knowledge management applied to the domain of Ecosystem Approach to Fisheries (EAF). According to FAO [5], EAF is an approach that:

> strives to balance diverse societal objectives, by taking into account the knowledge and uncertainties about biotic, abiotic, and human components of ecosystems and their interactions and applying an integrated approach to fisheries within ecologically meaningful boundaries.

We used this definition to create a conceptual model as schown in the UML diagram class of Figure 1.

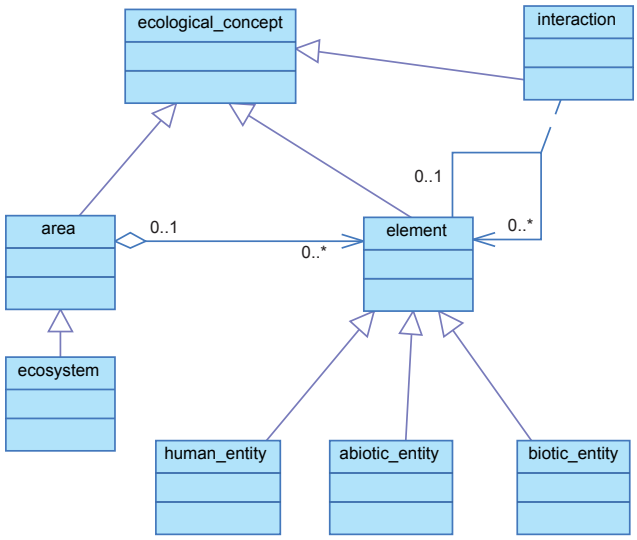

Fig. 1. UML class diagram translating Ecosystem Approach to Fisheries definition [5]

This UML diagram has been used to create a first set of top-level classes and properties for EAF domain (identifying real-world entities according to [2]): ecosystems (areas), their components as well as their interactions. The Figure 2 shows some RDF triples for the class *fish* instantiated for the species *exocoetus volitans*.

Similar objects have been created for hundreds of *species*, *fishing vessels* and *fishing gears*... which are all part of different *marine ecosystems*.

EAF requires thus the management of knowledge related to *marine ecosystems* and their *abiotic*, *biotic* and *human components*. However, this knowledge comes from different scientific studies driven by *researchers* which generate various *information resources*. These entities have to be described as well.

```
<ecosystems_def:fish rdf:about="&localfile;#exocoetus_volitans">
        <ecosystems_def:wormsId>http://www.marinespecies.org/aphia.php?p=taxdetails&amp;id=126385</
ecosystems_def:wormsId>
        <ecosystems_def:fishbaseId>http://www.fishbase.org/Summary/SpeciesSummary.php?id=1032</
ecosystems_def:fishbaseId>
        <ecosystems_def:faoId>EXV</ecosystems_def:faoId>
        <ecosystems_def:wikiId>http://commons.wikimedia.org/wiki/Exocoetus_volitans</
ecosystems_def:wikiId>
        <skos:prefLabel xml:lang="en">Flying fish</skos:prefLabel>
        <skos:prefLabel xml:lang="fr">Poisson volant</skos:prefLabel>
        <skos:prefLabel xml:lang="es">Pez volador</skos:prefLabel>
        <skos:altLabel xml:lang="fr">exocet</skos:altLabel>
        <skos:altLabel xml:lang="en">Tropical two-wing flyingfish</skos:altLabel>
        <skos:note xml:lang="fr">L'espèce Exocoetus volitans est étudiée au CRHMT en tant que
proie...</skos:note>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans1"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans2"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans3"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans4"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans5"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans6"/>
        <foaf:depiction rdf:resource="&ontologies;/resources#pictureExocoetusVolitans7"/>
        <ecosystems_def:is_prey_of rdf:resource="&localfile;#ale"/>
        <ecosystems_def:is_prey_of rdf:resource="&localfile;#bet"/>
        <ecosystems_def:is_prey_of rdf:resource="&localfile;#swo"/>
        <ecosystems_def:is_prey_of rdf:resource="&localfile;#yft"/>
        <ecosystems_def:used_data_source rdf:resource="&ontologies;/resources#dbStomac"/>
        <ecosystems_def:used_data_source rdf:resource="&ontologies;/resources#dbIsotopes"/>
        <geographic_objects_def:prefGeoGraphicObject rdf:resource="&ontologies;/
geographic_objects#wms_exocoetus_volitans"/>
    </ecosystems_def:fish>
```

Fig. 2. Examples of RDF triples summarizing our knowledge about species

2 Information resources related to EAF

Real-world entities of our domain (like species, fishing vessels, habitats. . .) are related to different kinds of information resources (pictures, spreadsheets, databases, satellites images, sensors data. . .)[15,16]. As illustrated in Figure 3, these entities are related as well to some people (agents) who are driving the scientific studies and generating the related information resources by running some processes.

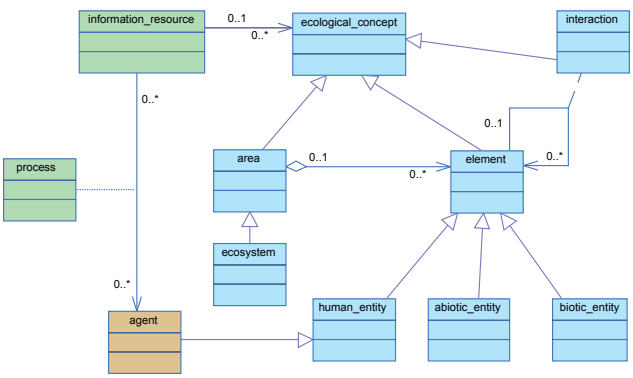

Fig. 3. Information resources, processes and agents related to entities of the domain

As many other laboratories working on ecological studies, the inventory of available data sources consists of various data types (which are subclasses of "information_resource" class in Figure 3):

– ecological observations from fieldwork (observers on-board fishing or scientific vessels collecting samples for data analysis: size, stomach content,

isotopes, contaminants, fatty acids...). These data are usually managed in spatial databases (Postgres / Postgis) or spreadsheets.
- satellites images to characterize environmental parameters of species habitats. These data are usually managed with binary formats (e.g. netCDF files for series of images).
- pictures that are collected by on-board observers or scientists,
- articles, reports... published by researchers,
- processes to run data analysis with different programming languages (R, IDL, Matlab...).

All these resources are usually described and managed with specific (meta-)data formats that impedes basic tasks like data discovery and retrieval. In addition to knowledge management, RDF is expected to facilitate data management (seamless access to metadata catalogues, codelist mapping...) by complying with widely used schemas.

3 Underlying standards

The description and the management of the information resources has to comply with well known standards to ensure that these resources will be made available for various communities of users. In particular, we target communities related to spatial, biodiversity and statistical information. In addition to information resources management, we selected as well existing standards to manage information about agents and domain entities (species...). An effort has been required to transform these (meta)data in RDF with EAF domain URIs.

3.1 Schemas for information resources and related agents

Many standards enable the description of information resources. Most of them consist in XML schema where keywords and mother metadata elements are described with literals (for species, characteristics, fishing vessels, agents... that are observed). This is the case with OGC metadata standards for spatial information (19115, 19119, 19110 [13]...), with metadata standards for biodiversity data (Ecological Metadata language / EML, TDWG standards like Darwin Core [18]...), with SDMX for statistical datasets [17]...

In order to describe and retrieve our information resources by using common metadata elements and URIs we decided to comply with following RDF schemas:

- DCMI [4] as metadata standard which can be used to describe any kind of information resource,
- dclite4g [19] Information model for metadata about geospatial data. ISO 19139 or EML metadata can be converted into dclite4g RDF metadata,
- SKOS [11] for description of terms and definitions related to information resources and concepts,
- FOAF [3] for description of agents (persons, institutes, projects) and their relationships,

– BIBO [7] for bibliographic references.

We aim to describe and make some of our data available as Linked Open Data by taking into account the 5 star development scheme for Open Data [2,1].

However, being able to describe information resources, processes and related agents requires URIs for ecological entities described in Figure 1.

3.2 Standards for domain entities

Among existing RDF schema relevant for our domain, we can mention:

– Previous work on ontologies for ecoinformatics [21].
– OBOE for modeling and representing scientific observations [9,10],
– Darwin Core [20] for sharing of information about biological diversity,
– FLOD [6] for fisheries domain.

These ontologies have been taken into account to map our ontology classes and properties as well as for raw data triplication.

3.3 RDF generation

We have two kinds of RDF triples which are generated from our information resources:

– statements instantiating our ontology with real-world entities (ecosystems and related components): species, fishing vessel, ecosystems... Each data source provides a set of entities (species, environmental parameters...) which are translated into instances of our ontology,
– RDF description of information resources, including related agents. More than basic descriptions, our goal here is to tag metadata with URIs of domain entities (previous item).

For now, we have been using an "ad hoc" approach to get RDF triples from each type of information resource as illustrated in Figure 4. Moreover in some cases, previous efforts can be reused:

– OGC metadata (ISO 19139) can be transformed in GENESI-DEC RDF metadata by reusing an existing XSL file from GENESI-DEC project,
– EML metadata can be transformed in OGC 19139 metadata with a GBIF XSL file,
– bibliographic references metadata can be exported in RDF (BIBO compliant) from Zotero (as well as references of pictures, videos if managed with Zotero).

In this case, the real challenge consists in adding some context to these RDF metadata by relating them to URIs of domain entities. This can be achieved by entity mining approach.

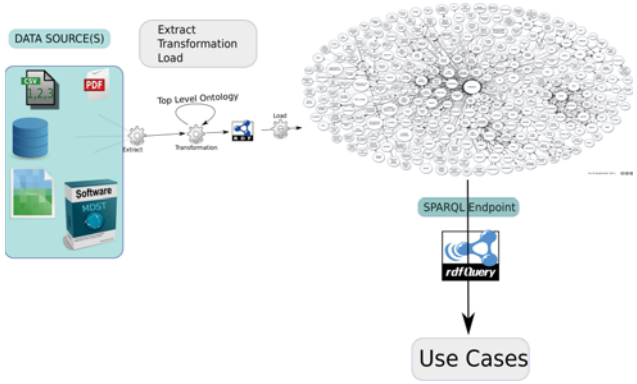

Fig. 4. Data sources to be triplified

4 RDF storage and server

Once classes and properties for EAF entities have been created and related objects (including information resources and agents) instantiated, these objects have been loaded with JENA in a triple store (Jena with TDB, preferred to Postgres for persistent storage and access) and have been made accessible through a SPARQL endpoint (the ontology is available as well at this URL: http://www.ecoscope.org/ontologies/main).

Another access has been set up to deliver search results through OpenSearch protocol with different data formats: HTML, RSS and RDF (Semantic extension).

That was needed to enable a set of use cases and make the content our knowledge base available online.

5 Related applications and products

In this section, we describe some use cases exploiting the content of the Ecoscope knowledge base. Our first use case was the setting up of a Web portal built on top of the ontology / knowledge base (through Jena for storage and access and Struts2 to set up Web pages on top of Jena). The goal was to satisfy basic use cases like data discovery and retrieval, knowledge summary about domain entities (species, fishing vessels. . .).

5.1 Metadata catalogue

One of the goal of using RDF metadata with URIs was to enable seamless access to different metadata catalogues without having to deal with specific standards. In particular OGC metadata (19115/39 used by INSPIRE), EML metadata (GBIF), Bibliographic references, Dublin Core metadata have been transformed to comply with a common set of metadata elements (cf section

3.1). Moreover these metadata are all anotated (and thus linked) with URIs of domain entities (cf section 3.2). This approach enables to query resources related to domain entities (e.g. yellowfin tuna) without having to restrict the search to specific standards, languages or terms. The search engine suggests all the results related to this entity and cluster the results according to their types Results can be related entities (e.g. preys of yellowfin tuna) or information resources like pictures, articles, databases, people... (see online application).

Fig. 5. Search engine for the Ecoscope knowledge base

More than inventorying existing information resources, our prior goal was to summarize available knowledge by themes / domain entities. This has been done by setting up fact sheets.

5.2 Fact sheets

The main purpose of our knowledge base was to feed the content of a Web portal by providing the knowledge about entities of interest for our laboratory (ecosystems, species, fishing vessels...). The main goal was the creation of fact sheets about these entities. To achieve this goal, A SPARQL query harvest all the triples related to a given entities and Jena objects are used by Struts2 to build some HTML views. Figure 6 gives an example of Web page for yellowfin tuna.

The fact sheets cluster related entities by type of relationships (*is predator of*, *is prey of*, *is exploited by*) and cluster related information resources by data types (pictures, spatial data and related processes / indicators).

Other visualization interfaces are available to present RDF triples as taxonomy or network.

Fig. 6. Fact sheet about Yellowfin tuna

5.3 Visualization of a food web

This use case illustrates what can be achieved with previous ontologies when applied to management of biodiversity data. The example of a food web is very relevant as it shows relationships between entities (species) that are either predators, preys or both. In Figure 7, we filtered RDF triples of the knowledge base to represent the food web related to tropical tuna trophic data (using prefuse API [8] for visualization of data). This application is interfaced with relational databases in order to enable users to spatially query the food web.

Fig. 7. Representation of a food web from RDF triples

The Figure 6 gives an example of Web page which is available online with other visualization application (e.g. Taxonomy).

5.4 Matching service

Another use case consists in using the Ecoscope knowledge base for the mapping of codification systems (code lists). For example, in EAF domain, species are often identified either by FAO codes in fisheries datasets or, most of the time, by Worms codification systems in ecological observations. This is an issue when researchers need to run some processes which require cross analyses between fisheries and ecological datasets. Indeed, in this case, there is a need to enable mapping between codes to merge datasets. A first application has been developed to enable such mapping at data export. We aim to deliver similar services in a generic way (in a programmatic way or through GUIs) to enable mapping between code lists of different schemas.

6 Outlooks

For now, RDF triples to link our data with entities and agents of the domain have been created in various ways. To make the knowledge base sustainable in the long term, we can't afford to feed it with ad hoc approaches. There is a need to harvest information from dedicated endpoints to facilitate updates by a workflow:

- databases and netCDF files will be turn in RDF through a single data server,
- pictures that are collected by on-board observers or scientists,
- articles, reports... published by researchers will be exported in RDF from Zotero Server,
- processes to run data analysis with different programming languages (R, IDL, Matlab...) will be described in RDF from OGC WPS metadata.

Another important improvement is related to the introduction of logical rules to infer some knowledge. A simple example consists in infering *competition* relationship between species from *predation* relationship. Indeed, two species are competing when they are predators of the same species (preys). This first use case is going to be implemented in the framework of iMarine FP7 program.

7 Conclusion

Our first application has demonstrated the interest of using ontologies and knowledge bases to satisfy different needs of researchers in our marine ecology laboratory: data discovery and retrieval, knowledge management and visualization (fact sheets, food web...). Such an application is worth but our current "ad hoc" approach still requires a lot of work to fill and update the database. To fix this issue, we aim to enable a RDF export from the Web portal which is used to access raw data (relationnal databases and netCDF files). We aim now to set up a workflow to facilitate RDF generation directly from the relational databases where ecological observations are stored and used by researchers to run scientific analysis. As a second step, we want ecological observations (raw data) to be

made available as well in RDF and to be linked with existing RDF triples (summarizing underlying knowledge: for example a RDF triple stating a predation relationship between two species should be related to hundreds of observations / facts proving it). The next goal consists in enabling RDF export from our main data sources, as already done with other standards (OGC, TDWG, SDMX).

References

1. Tim Berners-Lee. Linked data. July 2006.
2. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
3. Dan Brickley and Libby Miller. Foaf vocabulary specification. Namespace document, January 2010.
4. DCMI. Dublin core metadata initiative. *http://dublincore.org/*.
5. FAO. The ecosystem approach to fisheries. *FAO Guidelines for Responsible Fisheries*, (4):112, 2003.
6. FAO. Fisheries linked open data. *http://www.fao.org/figis/flod/*, 2011.
7. Frédérick Giasson and Bruce D'Arcus. Bibliographic ontology. Technical report.
8. Prefuse information visualization toolkit. http://prefuse.org/.
9. Joshua Madin, Shawn Bowers, Mark Schildhauer, and Matthew Jones. Advancing ecological research with ontologies. *Trends in Ecology & Evolution*, 23(3):159–168, March 2008.
10. Joshua Madin, Shawn Bowers, Mark Schildhauer, Sergeui Krivov, Deana Pennington, and Ferdinando Villa. An ontology for describing and synthesizing ecological observation data. *ECOLOGICAL INFORMATICS*, 2(3, Sp. Iss. SI):279–296, October 2007.
11. Alistair Miles and José R. Pérez-Agüera. Skos: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3):69–83, 2007.
12. Trina S. Myers, Ian Atkinson, and Ron Johnstone. Supporting coral reef ecosystems research through modelling re-usable ontologies. In *Knowledge Representation Ontology Workshop*, Proceedings of Conferences in Research and Practice in Information Technology (CRPIT), September 2008.
13. OGC. Open geospatial consortium, http://www.opengeospatial.org/.
14. Cynthia S Parr and Michael P Cummings. Data sharing in ecology and evolution. *Trends in Ecology & Evolution (Personal Edition)*, 20(7):362–363, July 2005. PMID: 16701396.
15. O. J. Reichman, Matthew B. Jones, and Mark P. Schildhauer. Challenges and Opportunities of Open Data in Ecology. *Science*, 331(6018):703–705, February 2011.
16. Leo Sauermann, Richard Cyganiak, and Max Völkel. Cool uris for the semantic web. Technical Memo TM-07-01, DFKI GmbH, Deutsches Forschungszentrum f ür Künstliche Intelligenz GmbH, February 2007. Written by 29.11.2006.
17. SDMX. Statistical data and metadata exchange. *http://sdmx.org/*, 2011.
18. TDWG. Taxonomic database working groug. *http://www.tdwg.org/*.
19. Jo Walsh and Pedro Goncalves. Dclite4g vocabulary. *http://dclite4g.xmlns.com/*.
20. John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin core: an evolving community-developed biodiversity data standard. *PLoS One*, 7(1):e29715, 2012.
21. Richard J. Williams, Nea D. Martinez, and Jennifer Goldbeck. Ontologies for ecoinformatics. *Journal of Web Semantics*, 4:237–242, 2006.

Detecting Semantic Overlap and Discovering Precedents in the Biodiversity Research Literature

Position Paper

Graeme Hirst*, Nadia Talent†, and Sara Scharf‡

*Department of Computer Science, University of Toronto
†Department of Natural History, Royal Ontario Museum
‡Independent Scholar
gh@cs.toronto.edu;nadia.talent@utoronto.ca;sara.scharf@gmail.com

Abstract. Scientific literature on biodiversity is longevous, but even when legacy publications are available online, researchers often fail to search it adequately or effectively for prior publications; consequently, new research may replicate, or fail to adequately take into account, previously published research. The mechanisms of the Semantic Web and methods developed in contemporary research in natural language processing could be used, in the near-term future, as the basis for a precedent-finding system that would take the text of an author's early draft (or a submitted manuscript) and find potentially related ideas in published work. Methods would include text-similarity metrics that take different terminologies, synonymy, paraphrase, discourse relations, and structure of argumentation into account.

Keywords: Biodiversity literature, taxonomy, systematics, natural language processing, Semantic Web, paraphrase, textual entailment, text similarity, discourse relations, structure of scientific papers.

1 Introduction

Scientific progress comes from building on, and occasionally overturning, past results. It is therefore a researcher's responsibility to know the history of the topic on which they are working, and this is so for two primary reasons: (1) to do the best possible work, building upon the state of the art, and neither duplicating what has already been done nor repeating the mistakes of the past; (2) to include in any publication of the work a literature review that allows the reader to understand the work in its broader context, compare it with cognate research, and evaluate it for quality and novelty. This requires the researcher both to maintain a knowledge of current research (*current awareness*) and to perform searches for relevant work in the legacy literature when their new work necessitates it (*finding precedents*).

Nonetheless, for a variety of reasons, researchers do not always adequately achieve these tasks, and this can lead to subsequent problems both for their own work and for that of other researchers. And this is particularly so in research in biodiversity, more

than perhaps most other sciences. Because of its longevous literature[1] and its need, in research on changes in biodiversity in ecosystems, to understand past conditions, finding precedents is both more important and more difficult than in the fast-moving don't-look-back-or-you'll-get-run-over sciences such as genomics.

In this position paper, we sketch the design of a proposed system that would draw on the mechanisms of the Semantic Web and methods in natural language processing to facilitate a search for precedents in the legacy biodiversity literature, especially (but not exclusively) the literature relating to systematics. It should be noted that what we are describing here is neither conventional search nor plagiarism detection (see footnote 8 below); our approach is influenced by research in the history of ideas in systematics on the detection of influence between authors and of independent re-invention (Scharf 2008).

2 What is a precedent, why do they matter, and why can they be hard to find?

We use the word *precedent* here, for want of a better term, to refer to any earlier published work or body of work that is, in an important way similar to, relevant to, or related to the current work in question. This is rather vague and subjective, but we can make it a little more concrete thus: An earlier published work is a precedent for current work if it has affected, or *should* have affected, the course of the newer work. This could include relevant methodologies, earlier attempts to solve the same problem, and earlier results and data. The most serious examples would be earlier work that is essentially the same as the present work (the new work is an independent re-invention), and, in particular, when the earlier work demonstrates that the new work is doomed to failure. Of primary interest to us in this paper are precedents in biodiversity research that, if not known and taken into account, render the current work seriously incomplete or erroneous.

Biodiversity research depends heavily on the legacy literature, which is the key source of important information about former biodiversity, and which also contains the results of massively time-consuming research that is difficult to replicate. The legacy literature of biodiversity includes a large component that is taxonomic literature. Besides the primary descriptions of new taxa, a major component of the taxonomic literature is synoptic volumes such as field guides, floras and faunas, synonymies, and 'manuals', which give varying levels of detail about the taxa present in a geographic area, including newly described taxa, summaries of opinion about previously defined taxa, and amended circumscriptions and descriptions. Modern synoptic works also include species-occurrence databases and analyses of biodiversity.

[1] "Natural history scientists work in fragmented, highly distributed and parochial communities, each with domain specific requirements and methodologies [Scoble 2008]. Their output is heterogeneous, high volume and typically of low impact, but with a citation half-life that may run into centuries" (Smith et al. 2009). "The cited half-life of publications in taxonomy is longer than in any other scientific discipline, and the decay rate is longer than in any scientific discipline" (Moritz 2005). Unfortunately, we have been unable to identify the study that is the basis for Moritz's remark.

Taxonomic nomenclature is a component of systematics that functions as a gateway to much of the taxonomic literature. It involves the application of the sets of rules that are laid down in the codes of nomenclature (ICZN 1999; McNeill et al. 2012) and periodically updated, with most provisions retroactively in force. The nomenclatural rules determine how the correct name for each species (or taxon of a higher or lower rank) must be determined. The *principle of priority* enshrined in the nomenclature rules holds as far back as the mid–eighteenth century, and literature of that vintage may be required to discover which name is correct. The definition of a taxon is anchored by the type specimen and the circumscription may be expressed either as a list of characteristics or as a list of specimens that the author considers to fit within the definition of the taxon. The specimen list may be either a list of typical specimens, or may be chosen to illustrate the range of morphological variation (or, potentially, the range of DNA sequences seen). Subsequent authors may wish to add to or subtract from the circumscription: common cases are (1) that a specimen of the other sex or a different life stage (such as a larva) is found, or (2) that a specimen originally cited is found to belong to a different taxon.

A taxonomist who wishes to create a new definitive list of the species in an geographic area or in a taxonomic group (a new "revision") must therefore search the legacy literature to find previous work that lists species in the area, or describes new species that might or might not be relevant, that amends previous descriptions, and (crucially) that works out the relationships between new or previously known species. They will need to find, evaluate, and cite prior publications that merge or split species (taxa), reclassify them into different groups, or assign new names to previously described species (taxa). All name alterations need to be re-evaluated in light of the rules of nomenclature now in force, which in practice means that previously ignored literature may resurface and lead the literature search into new areas. The precedents that were assumed for a work, and even the literature that was deliberately ignored for a work, may be listed in a way that requires a considerable sophistication in text understanding, for example in a book preface (e.g., Bentham and Hooker's *Genera Plantarum*).

Because of what has been termed the "citation gap" in the biodiversity literature (Payne et al. 2012), the taxonomic literature is massively undercited, and "such unintended omissions are likely to result in the decline of the [taxonomic] disciplines upon which the synoptic analyses depend" (Payne et al. 2012: p. 1350). This has occurred because the rules of nomenclature are now considered arcane by many researchers, and complete ignorance of the rules is common, not only among authors in ecology and biological taxonomy,[2] but lately even among the editors of major journals.[3] Large databases are being developed that already reduce the need to check the older literature, but their coverage is far from complete (Reveal 2012). Because of their ignorance and misunderstanding of the rules of nomenclature, the legacy literature becomes incompre-

[2] Systematics was traditionally a significant component of university biology courses, but the courses that provide this fundamental training have almost disappeared (Garnock-Jones 2013), replaced by courses that deal solely with molecular phylogenetic analysis, which is just one component of systematics.

[3] For an example of editorial problems, see the discussion in Taxacom at `http://mailman.nhm.ku.edu/pipermail/taxacom/2004-December/045547.html` et seq.

hensible to ecologists and inaccessible for biodiversity studies. But the consequences of mistakes, including failure to understand the older literature, can thus be very serious.[4]

Moreover, these kinds of mistakes may have a personal cost for their authors. When nomenclatural or taxonomic changes are referred to in later works, even in brief summaries, they usually carry a pointer to the authors who made the original change. Therefore, publications that err in this regard, if not ignored completely, are likely to be cited in a way that makes their transgressions apparent, an embarrassment for both the authors and the journal editors. For example, a taxonomic name may appear with an annotation such as *nomen dubium, nomen invalidum,* or *nomen illegitimum*, which indicates that the original authors erred. A correction may be published by later authors (neo- or lectotypification). When synonyms are listed, the authors commonly point to where their opinion differs from that of earlier authors, for example, *Synonyms: Leptospermum flavescens sensu W.L. Wagner et al. p.p., non Sm.* means that W.L. Wagner et al. included in the definition of *Leptospermum flavescens* some plants (*p.p. = pro parte* 'in part')) that did not match Smith's original description (*non Sm.*), and the present authors consider them to belong in another species; such a list may include implicit allegations that mistakes were made.

In the past, the principal problem had been lack of access to the required literature, but this is reducing, in large part due to the freely accessible Biodiversity Heritage Library[5] (Gwinn and Rinaldo 2009) and the (pay-walled) JSTOR collection, though much still remains inaccessible. But access helps only if researchers are willing to search this literature and can do so effectively. Non-technical barriers to doing so, in addition to the ignorance of the need and of the rules of nomenclature mentioned above, include time pressure, and the "Google effect" of just searching the Web and ignoring all but the top few results.

But even competent and well-intentioned researchers often have difficulties searching this literature. Simple Google-style keyword searches are frequently insufficient,[6] because in this literature, more so perhaps than most other fields of science, related concepts are often described or explained in different terms, or in completely different conceptual frameworks, from those of contemporary research. As a result, interesting and beneficial relations with legacy publications, or even with whole literatures, may remain hidden to term-based methods. In the case of taxonomy in particular, this implies the existence of what Nic Lughadha (2004) has called "hidden synonymies". The problem is compounded by ubiquitous Latin, non-obvious (to the modern reader) abbreviations, particularly Latin abbreviations and varied abbreviations of people's names, compact tabulations, and misspellings and multiple spellings of the same name.

[4] "International conventions and national or regional legislation concerning threatened or endangered animals specify the species or subspecies name of the animals that the law intends to protect. Thereafter, protection goes with the name rather than the endangered species itself. Any subsequent change in name could therefore affect conservation measures. The Commission often acts to protect the names of endangered species." — From the web site of the International Commission on Zoological Nomenclature (`http://iczn.org/content/conservation`)

[5] `http://www.biodiversitylibrary.org`

[6] Moreover, the quality of the OCR of many scans in the Biodiversity Heritage Library is presently so poor that keyword searches frequently result in false negatives.

Of course, none of this is to say that exact keyword matches are irrelevant or unhelpful. Term overlap can play its usual roles, and matches to names of taxa and of geographic locations are of particular importance.[7] However, our goal in the present work is to use semantic and structural relationships to discover the covert legacy literature that is not found with just a Google search or similar.

3 Foundational research

Ironically, we had great difficulty finding legacy literature on the topic of the difficulty of finding legacy literature, and on the topic of how researchers, in practice, search for and use this literature and the extent to which they do so.

The body of work that is perhaps most related to the former point is that of Swanson and colleagues (e.g., Swanson 1986; 1988; 1990) on identifying undiscovered public knowledge by analyzing the complementary but disjoint literature in two distinct fields of research and connecting knowledge in each to create new knowledge. For example, Swanson showed (1990; 1993) that studies on magnesium and studies on migraine, in two different fields, had terms in common, and the discovery that the two were related led in turn to the discovery that magnesium deficiency is connected with migraine. Superficially, the aim of this kind of analysis is the exact opposite of ours — it is looking at cases where, *a priori*, the authors are working in different research fields (rather than the same or closely related fields), and it does not operate at the level of the individual research paper. But methodologically it is similar nonetheless in that it is looking for an overlap or similarity in some aspect or aspects of the research. However, this work is limited in that the identification of related sub-fields was based simply on common terms used in both studies, and as we noted above, identical terminology cannot be assumed, even within a single research field. Moreover, the work needs, by its own background assumptions, to look at all possible pairings of topics of scholarship, and hence is prohibitively combinatorially explosive; in practice, a human must choose one topic or question as a starting point (Swanson 1993).

By contrast, in the approach that we will describe below, the search is constrained by assumption to a single, but large, field. This limits it sufficiently that it is computationally feasible with contemporary computing clusters. In the future, it will surely become computationally feasible to use our approach for Swanson's purposes.

4 Finding precedents in taxonomy and systematics

The confluence of research in natural language processing with Semantic Web technologies suggests the possibility in the near-term future of developing systems that would markedly improve researchers' ability to search and use the legacy literature in taxonomy and systematics. We assume the online availability of the literature itself — that is the continuing development of the Biodiversity Heritage Library (with improved

[7] A barrier that remains beyond the scope of this paper is the need for translation of literature written in languages not spoken by the searcher. Except for the special case of Latin, we do not address cross-lingual issues.

OCR), and access to the more-recent (still-in-copyright) twentieth-century literature in JSTOR and elsewhere. In this context, a precedent-finding system would take the text of an author's early draft (or a submitted manuscript) and find potentially related ideas in previously published work, matching not just words and phrases but ideas, regardless of how they are expressed. It would integrate current and expected near-term future research on the NLP technologies that we will describe below.[8]

We do not expect such a system to have a very high precision — many or most of its matches would be false alarms, although the design would attempt to minimize that. But the emphasis would be on high recall, bringing the potential matches to the attention of the user.

In the following subsections, we look at some of the primary elements, beyond literal keyword matching, of finding a match between new text and a potential precedent publication. We do not attempt a formal functional specification, which is the next step for this research, nor in the space available can we present examples, which would be textually large. We assume, without further comment, that a component for reasonably accurate translation of the Latin of taxonomic descriptions is available, and that the Latin is retained for keyword matching while the translation is used by other matching processes. We also assume that we have a component for recognizing taxonomic names in text, such as that of Koning, Sarkar, and Moritz (2005).

4.1 Paraphrase and similarity of meaning

The first element is the identification of sentences and phrases that are close in meaning. This has become an important research topic in computational linguistics in the last decade. It takes three forms; the first two are these:

1. *Paraphrase recognition*: identifying that two sentences or phrases are semantically equivalent or close to equivalent, even if very different in expression.
2. More generally, *recognizing textual entailment* (*RTE*): determining that the meaning of one sentence is entailed by, or is a consequence of, that of another. (Sentence-level paraphrase, then, can be thought of as mutual textual entailment.)

Dagan et al. (2013) provide a comprehensive survey of the techniques that have been developed for paraphrase recognition and RTE. Clearly, if we found this kind of a relationship between new work and a legacy publication, we would want to look further to see whether the latter might be a precedent.

The third form is this:

[8] Although there has been much research recently on *plagiarism detection* (see, for example, the evaluation lab overview by Potthast et al. (2012)), it is only peripherally relevant here, as it focuses primarily on finding matches for fragments of text that are precisely identical or differing in relatively minor ways, as when a plagiarizing student makes small changes in an attempt to evade detection. These are not the kinds of matches we are looking for. Current research in plagiarism detection has begun to take greater amounts of rewriting (including translation) into account (e.g., Barrón-Cedeño et al. 2013), making the task more like paraphrase detection (see below).

3. Measuring *semantic text similarity* (*STS*): identifying the degree to which two sentences, even if not paraphrases or entailing, are related in meaning.

Here, we are not looking for full equivalence or entailment, but rather trying to determine a degree of similarity or relatedness in meaning, and the methods that are used are rather different. Agirre et al. (2012) summarize the varied techniques and performance in a competitive evaluation of 35 STS systems. Even in the absence of equivalence or entailment, a high degree of relatedness throughout the two texts could indicate a potential precedent.

We expect that precedent-finding systems would draw on all three forms of this research. However, it should be noted that this research is presently limited to comparisons of pairs of sentences, whereas our goal inclues far broader comparisons long segments or complete texts, to find these relationships. So it will be important for this research to develop in this direction.

4.2 The low-level structure of scientific papers

The next element is the automatic analysis of the structure of scholarly discourse, especially scientific papers. Over the last decade, this has grown to become an important area of natural language processing (e.g., Ananiadou et al. 2012). This work endeavours to determine the structural purpose and discourse function of both individual sentences and of larger fragments of text in a scientific paper. Purposes or functions include such things as stating a claim, describing a gap in knowledge, criticizing or praising past work, and asserting the novelty of the present work (e.g., Teufel and Kan 2011; Angrosh et al. 2013a). This research also attempts to determine the purpose and scope of each citation in a paper (e.g., Siddharthan and Teufel 2007).

As this work becomes better and more mature, it can start to inform research on various relationships between texts (section 4.1 above), as the kind of information that it derives will be important in determining precedents. For example, if it is found that two sentences in different papers that are related in meaning are both claims, or both are statements of results, then we have a rather different situation with regards to identifying a precedent than if the sentence in the earlier paper is a result and the one in the later paper is a statement of the present state of the art.

The analysis of the structure of scientific texts will become more sophisticated in the future as it starts to incorporate more-detailed analysis of the discourse and rhetorical structures of text (e.g., Feng and Hirst 2012) — that is, the ability to find semantic discourse relationships between the clauses or sentences of a text, and then, in turn, the relationships that are built between larger fragments of text. That means not just the similarity or entailment relationships of section 4.1, but relationships such as CAUSE, CONTRAST, ELABORATION, and so on. And, in particular, it means finding them even when the author has left them only implicit in the text, which authors frequently do; in many contexts, human readers are able to recognize these relations without explicit textual cues, and authors tend to take advantage of this. Recognizing such implicit relationships is a current topic of research (Lin, Kan, and Ng 2009; Feng and Hirst 2013).

4.3 The argumentation structure of scientific papers

Our final element also relates to the structure of scientific papers, but at a higher level than the discourse relations. Ultimately, we would like to derive the structure of the overall argumentation[9] of a scientific text, and use that information too as a component of the matching process in our precedent-finding system. This is very difficult, even for people; a more realistic near-term goal based on current research (e.g., Lin, Kan, and Ng 2009; Feng and Hirst 2011) is to classify sentences as to their local role in the argumentation (e.g., premise, evidence) and use this information, and other identified discourse relations, to recognize larger components of the argumentation of the text and the kinds of argumentation scheme that it is using — for example, argument by analogy, or by induction, or by appeal to authority.

This could then allow matching of papers on the basis of the structure of the argumentation and how the content relates to this structure — or, indeed, independently of the content.[10] This kind of matching is less of an issue for the primarily fact-gathering aspects of searching the legacy literature that we described in section 2 above, but it would be of help in many other aspects of biodiversity (and other scientific) research.

4.4 Practical realization

Last, how would all this be realized in practice? Each item in the biodiversity and systematics legacy literature will need to be analyzed (including newly added items as they are published and as scanning of old literature continues) and annotated with an extensive representation for meaning and structure at all the levels of analysis. An important aspect of the representation and indexing of the legacy publications is that it must facilitate the process of checking for matches against new text, and must make this complex process as cheap as possible.

We anticipate that this representation would be based on XML and ontologies that are the topics of present-day research on mechanisms and resources for the Semantic Web. The annotation of some levels of analysis will be straightforward, such as the extraction of technical terms. Others will require further research and other design choices, as the nature of the representation will depend in part on the technical aspects of the methods chosen. For example, Dagan et al. (2013) list five distinct classes of methods for recognizing textual entailment; each implies different choices in the representation of the legacy text. One choice might involve annotating the text with details of the filled semantic roles of each sentence (Palmer, Gildea, and Xue 2010); another (not mutually exclusive) choice could be explicit annotation with contextually appropriate synonyms.

Practicality thus depends not only on our restriction of the domain (compared to the combinatorial problems of Swanson's approach, in section 3 above), but also on developing an effective representation.

[9] We refer, somewhat hyper-correctly, to *argumentation structure* to prevent the misinterpretation that we are talking about *argument structure* in the sense used in sentence-level syntax. We nonetheless refer to kinds of *argument* where there can be no terminological ambiguity.

[10] Retrieval of precedents by argumentation structure, without regard to the facts of any individual case, is also of particular concern to legal researchers (Dick 1991).

4.5 What's not included

The attentive reader will have observed that there are two things omitted from our proposal that might have been expected. The first is the use of citations and citation chains. One of our assumptions here is that our system is looking for things that are or might be completely disconnected, with respect to citations, from its starting point. Therefore, citations can play only a supporting role. Nonetheless, citations, including indirect connections, could still be a helpful factor in finding precedents; elaborating on this point is beyond the scope of this paper.

The other omission is semantic interpretation into a logical form, represented in XML, that draws on ontologies in the style of the original Berners-Lee, Hendler, and Lassila (2001) proposal for the Semantic Web. The problem with logical-form representation is that it implies a degree of precision in meaning that is not appropriate for the kind of matching we are proposing here. This is not to say that logical forms would be useless. On the contrary, they are employed by some approaches to paraphrase and textual entailment (section 4.1 above) and hence might appear in the system if only for that reason; but even so, they would form only one component of a broader and somewhat looser kind of semantic representation.

5 Conclusion

The precedent-finding system as we have sketched it here would be the culmination of a number of threads of research in computational linguistics and natural language processing and in document processing for the Semantic Web, and it can be thought of as a grand challenge for these fields. Moreover, we argue that by restricting our goals to the special case of the literature of systematic taxonomy and ecosystem biodiversity, we can achieve useful results in the near-term. But more generally, in a world in which increasingly interdisciplinary scholars must search an increasingly large legacy literature, precedent-finding systems would have great utility.

Acknowledgments. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada and the Canadian Newt and Eft Foundation. We are grateful to Heike Zinsmeister for helpful discussions.

Bibliography

Angrosh, M.A.; Cranefield, Stephen; Stanger, Nigel (2013a). Context identification of sentences in research articles: Towards developing intelligent tools for the research community. *Natural Language Engineering*, to appear.

Angrosh, M.A.; Cranefield, Stephen; Stanger, Nigel (2013b). Contextual information retrieval in research articles: Semantic publishing tools for the research community. *Semantic Web Journal*, to appear. `http://iospress.metapress.com/content/q7j3606047461315`

Agirre, Eneko; Cer, Daniel; Diab, Mona; Gonzalez-Agirre, Aitor (2012). SemEval-2012 Task 6: A pilot on semantic textual similarity. *First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, 385–393.

Ananiadou, Sophia; van den Bosch, Antal; Sándor, Ágnes; Shatkay, Hagit; de Waard, Anita (editors) (2012). *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Korea. http://aclweb.org/anthology-new/W/W12/W12-43.pdf

Barrón-Cedeño, Alberto; Vila, Marta; Martí, M. Antònia; Rosso, Paolo (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, to appear.

Berners-Lee, Tim; Hendler, James; and Lassila, Ora (2001). The Semantic Web. *Scientific American*, 284(5), May 2001, 34–43.

Dagan, Ido; Roth, Dan; Sammons, Mark; Zanzotto, Fabio Massimo (2013). *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool Publishers.

Dick, Judith (1991). Representation of legal text for conceptual retrieval. *Proceedings, Third International Conference on Artificial Intelligence and Law*, Oxford, 244–252. http://ftp.cs.toronto.edu/pub/gh/Dick-1991.pdf

Feng, Vanessa Wei and Hirst, Graeme (2011). Classifying arguments by scheme. *Proceedings, 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, 978–996.

Feng, Vanessa Wei and Hirst, Graeme (2012). Text-level discourse parsing with rich linguistic features. *Proceedings, 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Korea, 60–68.

Feng, Vanessa Wei and Hirst, Graeme (2013). Removing deleterious information to improve recognition of implicit discourse relations. Submitted.

Garnock-Jones, Phil (2013). The citation gap and its effects on taxonomy. In Blog: Theobrominated, 5 February 2013. http://theobrominated.blogspot.co.uk/2013/02/the-citation-gap-and-its-effects-on.html

Gwinn, Nancy E. and Rinaldo, Constance (2009). The Biodiversity Heritage Library: Sharing biodiversity literature with the world. *IFLA Journal*, 35(1): 25–34.

International Commission on Zoological Nomenclature 1999. *International Code of Zoological Nomenclature*, fourth edition. http://www.nhm.ac.uk/hosted-sites/iczn/code

Koning, Drew; Sarkar, Indra Neil; Moritz, Thomas (2005). TaxonGrab: Extracting taxonomic names from text. *Biodiversity Informatics*, 2, 79–82.

Lin, Ziheng; Kan, Min-Yen; Ng, Hwee Tou (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, Singapore, pages 343–351.

Moritz, Tom (2005). "Macro-economic case for open access." Talk at Library and Laboratory: The Marriage of Research, Data and Taxonomic Literature, London, 5–6 February 2005. http://barcoding.si.edu/LibraryAndLaboratory.htm or http://barcoding.si.edu/LibraryAndLaboratory/3-11_Moritz.pdf

McNeill, J.; et 13 al. (2012). *International Code of Nomenclature for algae, fungi, and plants (Melbourne Code) adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011*. A.R.G. Gantner Verlag KG.

Nic Lughadha, Eimear (2004). Towards a working list of all known plant species. *Philosophical Transactions: Biological Sciences*, 359(no. 1444): Taxonomy for the Twenty-First Century (2004-04-29), 681–687. http://www.jstor.org/stable/4142261

Palmer, Martha; Gildea, Dan; Xue, Nianwen (2010). *Semantic Role Labeling*. Morgan & Claypool Publishers.

Payne, Jonathan L.; et 8 al. (2012). A lack of attribution: Closing the citation gap through a reform of citation and indexing practices. *Taxon* 61(6): 1349–1351. http://www.ingentaconnect.com/content/iapt/tax/2012/00000061/00000006/art00030

Potthast, Martin; et 11 al. (2012). Overview of the 4th International Competition on Plagiarism Detection. *Proceedings, PAN 2012 Lab: Uncovering Plagiarism, Authorship and Social Software Misuse*. In: Forner, Pamela; Karlgren, Jussi; and Womser-Hacker, Christa (editors), *CLEF 2012 Evaluation Labs and Workshop — Working Notes Papers*, Rome.

Reveal, James L. (2012). A divulgation of ignored or forgotten binomials. *Phytoneuron* 2012-28: 1–64. http://www.phytoneuron.net/PhytoN-Divulgation.pdf

Scharf, Sara (2008). Multiple independent inventions of a non-functional technology: Combinatorial descriptive names in botany, 1640–1830. *Spontaneous Generations*, 2(1):145–184. http://spontaneousgenerations.library.utoronto.ca/index.php/SpontaneousGenerations/article/view/3552

Scoble, Malcolm J. (2008). Networks and their role in e-taxonomy. In: Wheeler, Quentin D. (editor), *The New Taxonomy* New York: CRC Press, 19–31.

Siddharthan, Advaith and Teufel, Simone (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. *Proceedings, Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, 316–323.

Smith, Vincent S.; et 4 al. (2009). Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. *BMC Bioinformatics*, 10(Suppl 14):S6. http://www.biomedcentral.com/1471-2105/10/S14/S6

Swanson, Don R. (1986). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30, 7–18.

Swanson, Don R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31, 526–557.

Swanson, Don R. (1990). Somatomedin C and Arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, 33, 157–186.

Swanson, Don R. (1993). Intervening in the life cycles of scientific knowledge. *Library Trends*, 41(4), 606–631.

Teufel, Simone; Kan, Min-Yen (2011). Robust argumentative zoning for sensemaking in scholarly documents. In: Bernadi, Raffaella et 4 al. (editors) *Advanced Language Technologies for Digital Libraries*, Lecture Notes in Computer Science, Volume 6699, 154–170.

Sponsors & Support