A semantic web faceted search system for facilitating building of biodiversity and ecosystems services

Marie-Angélique Laporte[1], Isabelle Mougenot[2], Eric Garnier[3], Ulrike Stahl[1,4], Lutz Maicher[1,5], and Jens Kattge[1,4]

[1] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany,
[2] UMR 228 ESPACE-DEV, Maison de la Télédétection 34093 Montpellier, France,
[3] Centre d'Ecologie Fonctionnelle et Evolutive (UMR 5175), 1919 Route de Mende, 34293 Montpellier Cedex 5, France,
[4] Functional Biogeography Research Group, Max Planck Institute for Biogeochemistry, 07701 Jena, Germany,
marieangelique.laporte@gmail.com, isabelle.mougenot@ird.fr ,eric.garnier@cefe.cnrs.fr, lutz.maicher@moez.fraunhofer.de, ustahl,jkattge@bgc-jena.mpg.de

Abstract. To address biodiversity issues in ecology and to assess the consequences of ecosystem changes, large quantities of long-term observational data from multiple datasets need to be integrated and characterized in a unified way. Linked open data initiatives in ecology aim at promoting and sharing such observational data at the web-scale. Here we present a web infrastructure, named Thesauform, that fully exploits the key principles of the semantic web and associated key data standards in order to guide the scientific community of experts to collectively construct, manage, visualize and query a SKOS thesaurus. The study of a thesaurus dedicated to plant functional traits demonstrates the potential of this approach. A point of great interest is to provide each expert with the opportunity to generate new knowledge and to draw novel plausible conclusions from linked data sources. Consequently, it is required to consider both the scientific topic and the objects of interest for a community of expertise. The goal is to enable users to deal with a small number of familiar and conceptual dimensions, or in other terms, facets. In this regard, a faceted search system, based on SKOS collections and enabling thesaurus browsing according to each end-users requirements is expected to greatly enhance 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, cli-

matology, soil sciences, etc [1]. To address biodiversity issues, it is now widely accepted that a functional approach has strong potential. Over the last decades, trait-based research has generated huge volumes of data, within multiple contexts of observations and experiments [1]. These data sets can be obtained via very different study contexts and are often described in highly specialized terms. Numerous traits can be measured, for instance, on plants [2]. But data generated by functional ecology are only minimally reused or shared within the community, or over communities, mainly due to data heterogeneity [1]. 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.

Technologies developed under the Semantic Web initiative are particularly suitable for the sharing and the dissemination of information within a community of experts. SKOS (Simple Knowledge Organization System) [3] provides a common format to manage thesaurus adequately. The final purpose of a thesaurus is to facilitate the integration and the navigation of the information available in multiple data sources. Each SKOS thesaurus can be considered as a publicly available relevant resource on the web and can be enriched via meaningful navigation between thesauri. Linked Data initiatives put a strong emphasis on representing KOS (Knowledge Organization System) for both data discovery and data access [4]. The LOD initiative (Linked Open Data) are more and more adopted by a large panel of data providers and make publically available each day data from a wide range of disciplines including the Life Science field [5]. As a result, the LOD contains more than hundred datasets [6], which can be freely used in dozens of different contexts. The potential of each data is then fully exploited. In this regard we want to emphasize the critical importance of properly connecting observational data with each other. Defining new vocabularies based on the Semantic Web standards, as SKOS, makes them fully interoperable and allows to directly benefit from data already published in this form on the Web of Data.

In this paper, we present a complete system dedicated to the ecological community allowing it to create, manage, visualize and query a SKOS thesaurus. In the context of biodiversity studies, the TOP (Trait of Plants) thesaurus [7] is used to semantically annotate scientific data managed through heterogeneous data sources, such as the TRY database [8] or the Plant Ontology (PO) [9]. The TOP thesaurus is then exploited through a faceted search engine that reflects community interests and preferences, to facilitate the appropriation of the TOP thesaurus by various end-users. The facets act also as access points on the interrelated data sources in guiding their navigation. In this paper, we foccussed on how end-user points of view can be developed and implemented.

This article is organized as follow:

- Section 2 quickly introduces the approach driven with the Thesauform tool to build the TOP thesaurus as a collaborative product, and presents how the faceted search enhances the information retrieval in ecology and beyond this.

- Section 3 explains how the thesaurus is used for integration purposes. The TOP thesaurus aggregates data from disseminated datasources with the purpose of both enriching and facilitating data interpretation.

- Finally, section 4 summarizes and discusses the strengths of our approach and refers to future works.

2 Faceted search to improve information retrieval in ecology

In order to build a collective thesaurus, our previous work focused on the development of a tool, named Thesauform, dedicated to assist domain experts in this task. The Thesauform tool fully relies on semantic web standards, while providing a flexible and user-friendly environment for domain experts. Twenty different experts from the functional plant trait community has used the Thesauform tool to describe the different functional plant traits in use in the domain. 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 can be used as a bibliographic resource about plant traits information, since it is available as a web resource[5]. The TOP thesaurus fulfills its initial role to provide a standard vocabulary available to the functional ecology community, and extends beyond the basic needs to ease information retrieval. During the thesaurus building steps, many users complained about the hierarchical structure of the thesaurus, arguing that the concepts should be ordered in a different way, and even that the thesaurus should present different hierarchies. Indeed, in some cases, users were not able to find quickly and easily the concepts on which they wanted to work on. In this context, a system considering end-user points of view has been developed and offers a faceted search engine.

Classic semantic search engines based on controlled terms have been widely used to query data in the life science fields. For instance, Bioportal[6] is a web portal providing the interrogation of multiple ontologies or controlled vocabularies based on controlled terms. Although this kind of search mechanism offers a first control over the terms used for the search, it suffers from limitations since it can be difficult for an inexperienced end-user to find the relevant controlled terms to use [10]. Indeed, with classic semantic search engines, controlled terms are most of the time displayed through an auto-completed search field. This would suggest that the user has a prior knowledge of the content of the data model to

[5] http://trait_ontology.cefe.cnrs.fr:8080/Thesauform/vizIndex.jsp (developed as a proof of concepts)

[6] http://bioportal.bioontology.org/

query. Furthermore, information cannot always fit into a well-defined hierarchy that users know how to browse [11]. To overcome these limitations, a well known searching and filtering technique coming from the field of Information Retrieval, the faceted search, is widely used over the web.

The faceted search is an interesting solution as it facilitates the thesaurus appropriation by the end-users by helping users to define their search needs [12, 10]. In this context, facets will lead to translate the vague query that a user can have, to a precise query in the system. The MUMIA [7] web site gives a simple definition of faceted search (also called faceted navigation or faceted browsing). Faceted search is "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". In other terms, each facet typically corresponds to the common features shared by a set of objects. These features are used to filter the results. Finally, in such search engines, the user is guided (no dead-end query) as the results are filtered using relevant parameters or categories, each category reflecting both the need of users in the thesaurus navigation environment and structuring the information so others can find it.

In the TOP thesaurus, five facets have been defined according to users feedbacks with the first purpose of facilitating information retrieval. In thesaurus or in any other controlled vocabulary or ontology, concepts can be assembled into semantically meaningful groups that will correspond to facets. Since facets are closely linked to both thesaurus visualization and thesaurus restitution, and not to thesaurus structure or to the information it carries, existing good practices recommend to define facets as skos:collection [13], gathering concepts with common features. The use of skos:collection allows thus to combine concepts regarding a specific subject independently of the hierarchical classification of concepts in the concept scheme. An example of facets is described in Figure 1. The functional plant trait concept Specific Leaf Area (SLA) is classified under the concept of Morphology in the thesaurus, since this trait refers to the morphology of a plant. In Figure 1, SLA is grouped with the concepts Leaf Phenology and 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 considered as two access points to query the thesaurus by organizing the thesaurus in two different ways. Each user can then choose which access point to use to query the thesaurus according to his own preferences. The organization of the thesaurus concepts into different views corresponding to different hierarchies makes perfect sense during the user query, the reorganization of the thesaurus information facilitating the navigation in the thesaurus.

We conclude that a faceted search system is suitable to assist users in their information retrieval. Developing such a system based on facets allows to guide

[7] http://www.mumia-network.eu/index.php/working-groups/wg4

```
:OrganFacet    a skos:Collection;
       skos:member :Leaf;
       skos:member :Root;
       ...
       skos:member :Seed;
       skos:member :Flower .
:Leaf    a skos:Collection;
       skos:member :LeafArea;
       skos:member :Specific Leaf
       Area;
       ...
       skos:member :LeafLifespan .
```

```
:SizeFacet    a skos:Collection;
       skos:member :Area;
       skos:member :Length;
       skos:member :Density;
       skos:member :Mass;
       skos:member :Volume.
:Area    a skos:Collection;
       skos:member :LeafArea;
       skos:member :Specific Leaf
       Area;
       ...
       skos:member :XylemArea .
```

Fig. 1. Example of facets represented in Turtle format (RDF serialization format). Two facets are presented. The Organ facet allows to query the thesaurus using a plant organ. The Size facet is used to query the thesaurus according to the type of measure. The members of the facet values (i.e. :Leaf) come directly from the concept scheme hierarchy. The selection of Leaf from the facet Organ selects only traits measured on Leaf (belonging to the skos:Collection Leaf). Then, by selecting Area from the Size facet, the results are refreshed to contain only traits measured on Leaf and measuring an Area: LeafLifespan and XylemArea are then deleted from the results list.

the consultation of datasets in an intuitive way for the user. As the TOP thesarus is used as an access point to disseminate information, data sources semantically annotated with its concepts will be able to benefit from faceted search engines as well.

3 Facets for facilitating the access to disseminated data

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 great advantage of both enriching and facilitating the data interpretation, which requires information from different domains. Consequently, TOP thesaurus concepts have been linked to two different data sources, the TRY database, the biggest functional plant traits database, and the Plant Ontology (PO), the reference controlled vocabulary describing plant entities. A real advantage of SKOS is to provide properties dedicated to the establishment of cross-references between thesauri. The mappings between the TOP thesaurus and TRY and PO relies on SKOS properties dedicated on this purpose: the exactMatch and relatedMatch properties. The mappings to both TRY and PO have been managed automatically, based on term similarity. However, for TRY, the proposed mappings have been then manually curated by an expert in order to be validated. For TRY, only the exact matches has been saved in the mapping file. For PO, as plant entities are not traits, the matches have been recorded as related matches in the mapping file using the relatedMatch property.

The benefit of linking TOP thesaurus concepts to TRY is twofold. First, the mapping TOP/TRY allows to unify the access to TRY data, managing the terms' heterogeneity used to describe TRY data. The TRY database can then take full advantage of the different semantic search engines set up to query the TOP thesaurus information. Secondly, such a mapping will enrich and complete the information of the thesaurus itself by adding meta-information coming directly from the TRY database. For instance, on the given trait information webpage, in addition of the trait information themselves (preferred term, definition, synonyms,..), the TRY observation number, the geo-referenced observation number, the number of different species on which the given trait has been measured are also displayed. This information can be useful for the user that will be then able to get indications on the community interest for the given trait and the number of available data on that specific trait.

The mappings established between the TOP concepts and the PO concepts allow assigning a reference for the plant entities cited in most of TOP trait definitions. For instance, the definition of the Specific Leaf Area does not need then to explicit what is actually a leaf. This part of the definition is provided by the PO mapping as PO clearly defined what is a leaf. Moreover, such a mapping approach will be highly beneficial to link data used in ecology or agronomy to data used in genomics following a Linked Data approach. As the TOP thesaurus is mainly used by the ecology community and PO is mainly used in the genomic field, the mapping established between PO and TOP 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 resulting semantic web-infrastructure is displayed in Fig. 2. The TOP thesaurus addresses the need of organizing the available information in a unifying way in a context of biodiversity studies. Indeed, the thesaurus aggregates data from different data sources in order to build new biodiversity models where the faceted system tied to the thesaurus plays a great role. Facets can be now used to access information from these relevant and disseminated data through the thesaurus, with a reorganization of the information. In fact, TRY and PO are queried through the facets and no more through the original way the information was structured. The ecological community has therefore a full-integrated access to disseminated sources in a way that reflects their interests, thus facilitating both their discovery and their reuse.

The trait information coming from the TOP thesaurus will be mainly accessed by experts from the ecology domain. Considering this, as a proof of concepts, we based our work on a user-friendly and easy to use interface, to assist experts in their access and retrieval of pertinent information. We implemented 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. A unique aspect of our work is the implementation of a faceted search engine based on skos collections. This enhances the semantic search of trait by providing the opportunity to the user to choose his own filters.

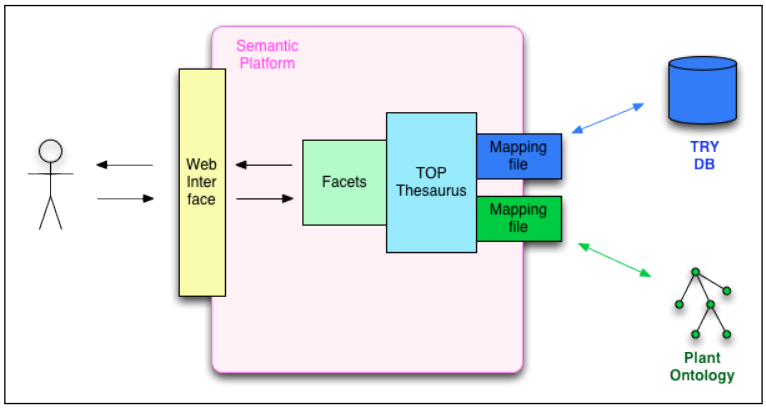

Fig. 2. Unifying system of plant trait modelling, based on Semantic Web technologies. The TOP thesaurus aggregates data from the TRY database and the Plant Ontology with the purpose of both enriching and facilitating data interpretation. The faceted search system is used to query the linked data and filter the results according to users preferences.

4 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 [14, 15]. 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 and assists any community of experts to manage their datasets, and to interconnect them with data and data standards from related communities. In this regard we have emphasized the critical importance of properly connecting observational data with each other. The construction of a SKOS vocabulary facilitates the definition of clear semantic bridges between different data sources. The participation of experts, not only for the construction of thesauri, but also to validate the work related to semantic annotations, strengthens the proposed approach.

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. First, facets ensure flexibility by playing an assistance role to users by reorganizing the thesaurus terms into meaningful groups. The faceted system supports users to specify their queries and then to drill down to results. Second, data sources semantically annotated with concepts coming from the TOP thesaurus can benefit from faceted search engine traits as well. As a consequence, facets are be used to discover and access disseminated information from heterogeneous data sources. A next step will be to propose mappings to more external resources, numerous relevant ontologies can be found on the NCBO BioPortal website. The approach championed in this paper has been to

base our work on the continuity of the Open Linked Data initiative, based on Semantic Web techniques. Future work will be focused on how these facets could be automatically built from both existing literature and ontologies.

References

1. Michener, W.K., Jones, M.B.: Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution **27**(2) (February 2012) 85–93
2. 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)
3. Isaac, A., Summers, E.: SKOS Simple Knowledge Organization System Primer. W3C Technical Report (2008)
4. Panzer, M., Zeng, M.L.: Modeling classification systems in skos: some challenges and best-practice recommendations. In: International Conference on Dublin Core and Metadata Applications. (2009) pp–3
5. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics **41**(5) (2008) 706 – 716 Semantic Mashup of Biomedical Data.
6. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology **1**(1) (2011) 1–136
7. Laporte, M.A., Garnier, E., et al.: Thesauform—traits: A web based collaborative tool to develop a thesaurus for plant functional diversity research. Ecological Informatics **11** (2012) 34–44
8. Kattge, J., DÍaz, S., Lavorel, S., Prentice, I., Leadley, P., & al.: TRY - a global database of plant traits. Global Change Biology **17** (2011)
9. Walls, R., Cooper, L., Elser, J., Stevenson, D.: The Plant Ontology: A Common Reference Ontology for Plants. wiki.plantontology.org (2010) 2010
10. Heim, P., Ertl, T., Ziegler, J.: Facet graphs: Complex semantic querying made easy. In: The Semantic Web: Research and Applications. Springer (2010) 288–302
11. Uddin, M.N., Janecek, P.: Faceted classification in web information architecture: A framework for using semantic web tools. Electronic Library, The **25**(2) (2007) 219–233
12. Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., ping Yee, K.: Finding the flow in web site search. Commun. ACM 2002
13. Brugman, H., Malaisé, V., Gazendam, L.: A web based general thesaurus browser to support indexing of television and radio programs. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). (2006) 1488–1491
14. Reichman, O.J., Jones, M.B., Schildhauer, M.P.: Challenges and Opportunities of Open Data in Ecology. Science **331**(6018) (February 2011) 703–705
15. 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)