Proceedings of the 2nd International Workshop on Semantics for Biodiversity (S4BioDiv 2017)

Alsayed Algergawy, Naouel Karam, Friederike Klan, Clement Jonquet

▶ To cite this version:

HAL Id: lirmm-01605671

https://hal-lirmm.ccsd.cnrs.fr/lirmm-01605671

Submitted on 17 Jan 2018

Preface

Biodiversity research aims at comprehending the totality and variability of organisms, their morphology, genetics, life history, habitats and geographical ranges. It usually refers to biological diversity at three levels: genetics, species, and ecology. Biodiversity is an outstanding domain that deals with heterogeneous datasets and concepts generated from a large number of disciplines in order to build a coherent picture of the extend of life on earth. The presence of such a myriad of data resources makes integrative biodiversity research increasingly important, but at the same time very challenging. It is severely strangled by the way data and information are made available and handled today. Semantic Web techniques have shown their potential to enhance data interoperability, discovery, and integration by providing common formats to achieve a formalized conceptual environment, but have not been widely applied to address open data management issues in the biodiversity domain.

The 2nd International Workshop on Semantics for Biodiversity (S4BioDiv) thus aimed to bring together computer scientists and biologists working on Semantic Web approaches for biodiversity and related areas such as agriculture or agro-ecology. The goal was to exchange experiences, build a state of the art of realizations and challenges and reuse and adapt solutions that have been proposed in other domains. The focus was on presenting challenging issues and solutions for the design of high quality biodiversity information systems based on Semantic Web techniques. The workshop was a full-day event on October 22nd co-located with the 16th International Semantic Web Conference (ISWC 2017), October 21-25, Vienna, Austria.

In total, 13 paper submissions presenting new research results and ongoing projects have been submitted. All of these were reviewed by at least three members of the program committee. Out of the submitted contributions, 6 full papers and 4 poster papers have been accepted for presentation at the workshop and publication in these proceedings.

The program included two keynote talks highlighting two vital and challenging topics related to biodiversity research and Open Science in general. Alison Specht, director of the Centre for the Synthesis and Analysis of Biodiversity (CESAB), talked about "Engaging the Domain Expert: Is it just a Dream?". Oscar Corcho, full professor at the Ontology Engineering Group, ETSI Informáticos, Universidad Politécnica de Madrid, Spain presented his thoughts "Towards Reproducible Science: A few Building Blocks from my Personal Experience". To stimulate interdisciplinary debate, the workshop encompassed a one-hour panel discussing controversial topics in the field.

We would like to thank the ISWC workshop chairs Aidan Hogan and Valentina Presutti for their kind support. We are also grateful to the workshop's program committee consisting of

Birgitta König-Ries (Friedrich-Schiller-Universität Jena, Germany)

Alsayed Algergawy, Naouel Karam, Friederike Klan & Clement Jonquet
S4BioDiv Chairs
September 2017

Engaging the Domain Expert: Is It just a Dream?

Alison Specht

CEntre for the Synthesis and Analysis of Biodiversity (CESAB), Aix-en-Provence, France
alison.specht@fondationbiodiversite.fr

Abstract

Much of the work of creating open data repositories, and of enabling the sharing and discovery of their contents, is conducted by informatics specialists on behalf of the communities that they serve. Some, such as the genomics, remote sensing and medical communities, have a good record of participation in data sharing, custodianship, discovery and re-use, and are familiar with the procedures involved. Ecological scientists, especially those engaged in experimental and observational studies where the data are gathered personally, not through automatized means, are not so participative.

Ecological scientists are, for a start, very possessive of their data; they are collected through their own blood, sweat, and tears. They know well the value of long-term observations and experiments in order to make scientific decisions about change (extinctions, rarity and so on), and are among the first to support their need, but they are reluctant to share their own data. In addition, the big questions that face ecological scientists demand that they work across disciplines as well as collaborate within their own community. The rates of data sharing and re-use by ecologists are among the lowest in the sciences, whereas data loss is comparatively high. Why? How can we improve this situation?

I will discuss the particular challenges of dragging ecologists into the open data world. I will comment on the steps in the research data lifecycle where the requirements of participation are particularly distracting for ecologists. I will describe an initiative that has been developed to enable domain specialists to take control of their terminological understanding in a manner that facilitates the assembly of heterogeneous data sets, and at the same time prepares them for open delivery of their compiled data. I suggest that through this process, they will better understand the requirements for data sharing and custodianship, and will be more likely to participate in all facets of the open data world, improving the availability of long-term data sets!

Biography

Alison Specht is an environmental scientist with interest in facilitating trans-disciplinary, convergent research between scientists, policy-makers and managers

to improve environmental outcomes, and in improving data management and preservation of archival data for effective long-term monitoring.

From 2009 to 2014 she was the director of the Australian Centre for Ecological Analysis and Synthesis (www.aceas.org.au), a facility of the Terrestrial Ecosystem Research Network (www.tern.org.au), the first synthesis centre in the Southern Hemisphere. Since September 2015 she has been Director of CESAB, the CEntre for the Synthesis and Analysis of Biodiversity in France (cesab.org). She is a core partner of the International Synthesis Consortium (synthesis-consortium.org). She has been a member of the DataONE (www.dataone.org) Usability and Assessment Working Group since its inception in 2010.

Towards Reproducible Science: A Few Building Blocks from my Personal Experience

Oscar Corcho

Ontology Engineering Group, ETSI Informáticos, Universidad Politécnica de Madrid, Spain
`ocorcho@fi.upm.es`

Abstract

It is well understood that achieving Reproducible Science across all scientific disciplines is an extremely ambitious goal that will be really difficult to achieve. However, as far as it could be, there are many small steps that can be taken towards improving our way of doing, communicating and advancing Science, by making the experiments that we describe in our scientific papers easier to reproduce.

In this talk, I will talk about some of the efforts that we have been working on in the context of our research group, focused on achieving a more Reproducible Science.

First, our work on ontologies for the representation of wetlab laboratory protocols (for plant genomics). We have been working for a few years on analysing manually papers describing laboratory protocols, deriving a representation for them, understanding how Instruments, Reagents, Outputs, etc., have to be identified and annotated, and working on an annotation tool for those creating lab protocols. Finally, we are now in the process of publishing this laboratory protocols as Linked Data. All this work is also related to other works that we have been doing in the past in collaboration with other institutions for the description of research objects and for the description of scientific in-silico workflows.

Second, the work that we are doing in the context of the STARS4ALL EU project, where we are trying to provide support to the research (and activists) community working on light pollution and the negative effects of artificial light at night. More specifically, we are working on making research data available as open data, including the deployment of a research data hub for the community, as well as creating ontologies that can be used by public institutions in order to release data about public lighting.

Finally, I will discuss on what I believe that is still needed in order to achieve the broader goal of Reproducible Science and will open a discussion on the current barriers to achieve this goal.

Biography

Oscar Corcho is Full Professor at Departamento de Inteligencia Artificial (Facultad de Informtica , Universidad Politcnica de Madrid), and he belongs to the Ontology Engineering Group.

His research activities are focused on Semantic e-Science and Real World Internet, although he also works in the more general areas of Semantic Web and Ontological Engineering. In these areas, he has participated in a number of EU projects (DrInventor, Wf4Ever, PlanetData, SemsorGrid4Env, ADMIRE, OntoGrid, Esperonto, Knowledge Web and OntoWeb), and Spanish R&D projects (CENITS mIO!, Espaa Virtual and Buscamedia, myBigData, GeoBuddies), and has also participated in privately-funded projects like ICPS (International Classification of Patient Safety), funded by the World Health Organisation, and HALO, funded by Vulcan Inc.

Previously, he worked as a Marie Curie research fellow at the University of Manchester, and was a research manager at iSOCO. He holds a degree in Computer Science, an MSc in Software Engineering and a PhD in Computational Science and Artificial Intelligence from UPM. He was awarded the Third National Award by the Spanish Ministry of Education in 2001.

He has published several books, from which Ontological Engineering can be highlighted as it is being used as a reference book in a good number of university lectures worldwide, and more than 100 papers in journals, conferences and workshops. He usually participates in the organisation or in the programme committees of relevant international conferences and workshops.

AnnoSys2: Reaching out to the Semantic Web

Okka Tschöpe[1], Lutz Suhrbier[2], Anton Güntsch[3] and Walter G. Berendsohn[4]

BGBM, Freie Universität Berlin, Germany
[1] o.tschoepe@bgbm.org
[2] l.suhrbier@bgbm.org
[3] a.guentsch@bgbm.org
[4] w.berendsohn@bgbm.org

Abstract. AnnoSys is a web-based open-source system for correcting and enriching specimen data in publicly available data portals, thereby bringing traditional annotation workflows for biodiversity data to the Internet. During its first phase, the project developed a fully functional prototype of an annotation data repository for complex and cross-linked XML-standardized data, including back-end server functionality, web services and an on-line user interface. Annotation data are stored using the Open Annotation Data Model and an RDF-database. The current project phase aims at extending the generic qualities of AnnoSys to further structured data formats including RDF data with machine readable semantic concepts, thus opening up the data gathered through AnnoSys for the Semantic Web. We developed a semantic concept-driven annotation management, including the specification of a selector concept for RDF data and a repository for original records extended to RDF and other formats. Since many of the biodiversity data standards in use are still not defined in a semantic-web compliant way, mechanisms for referencing elements in such data sets need to be developed. We therefore developed an AnnoSys ontology based on DwC RDF terms and the ABCD ontology, which deconstructs the ABCD XML-schema into individually addressable RDF-resources published via the TDWG Terms Wiki. We mapped the terms from these standards into annotation types we defined, based on semantic concepts.

Keywords: AnnoSys, Ontology, Annotation.

1   Introduction

Biodiversity data are aggregated, linked and made globally accessible via a range of Internet portals and services. Globally, natural history collections contain 2–3 billion specimens [1]. These provide materials and primary data for a wide range of research questions and form the basis for the classification of organisms into species and other "taxa". Traditionally, specimens are annotated by researchers with written annotation labels which are applied directly to the physical object, thus becoming accessible to succeeding observers of the specimen. These annotations improve the data quality of

the collection and document research developments over time (e.g. the understanding of taxon concepts).

To ensure the continuance of the traditional data sharing and incremental documentation of specimens in the on-line environment, the AnnoSys project developed an annotation data repository [2] for complex XML data following the ABCD [3] and DwC [4] standards. This includes back-end server functionality, web services and an on-line user interface [5]. Annotation data are stored using the Web Annotation Data Model [6] and an RDF-database [7].

In a second step, AnnoSys2 aims at extending the generic qualities of AnnoSys to further structured data formats including RDF data with machine readable explicit semantic concepts.

2 Motivation/State of the art

Since many of the biodiversity data standards in use are still not defined in a semantic-web compliant way, mechanisms for referencing elements in such data sets need to be developed. We therefore compiled an AnnoSys ontology based on DwC RDF terms and the ABCD ontology, which deconstructs the ABCD XML-schema into individually addressable RDF-resources published via the TDWG Terms Wiki [8].

One of our motivations for the new ontology was to harmonize annotatable elements to allow unambiguous comparability between different versions of a record. For example, depending on the data publishing portal, a record can be displayed either in the DwC or the ABCD standard. In AnnoSys 1 we were facing the problem that those records were not directly comparable, because not all ABCD elements are part of the DwC standard and vice versa. We therefore needed different versions histories for different data standards (DwC, ABCD 2.06, ABCD 2.1 etc.). The AnnoSys ontology defines matching rules describing how these different elements are transformed into annotatable elements, resulting in harmonized records with only one, unambiguously comparable version history.

Additionally, via the different SKOS-relations equivalence levels for matches of elements can be specified, which potentially allows restricting the use of elements to those with a minimum level of equivalence. This may be important for data formats that need to be integrated in the future

3 Model construction

We used Protégé [9] to build an AnnoSys ontology based on DwC terms [4] and the ABCD ontology [8], which uses ABCD property terms as RDF predicates. We created a subclass "RecordConcept" comprising all ontology concepts as a subclass of *skos:concept* (Fig. 1). We also defined nine different "annotation types" as instances of the SubClass "annotation type" of *oa:Motivation* (Fig.1, Fig. 2). Individual concepts were related to the different annotation types via the skos:related relation. We then mapped the elements of the two standards to semantic concepts using the

skos:Concepts exactMatch, broadMatch, narrowMatch, or closeMatch, respectively, to represent the different levels of matches (Table 1).
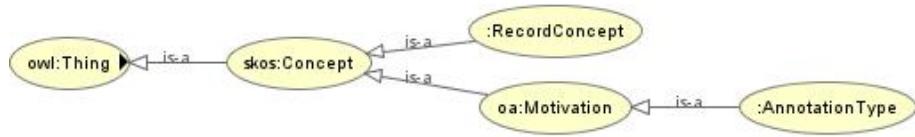
Fig. 1. Subclasses of skos:Concept in the AnnoSys Ontology.

Concepts that refer to identifiers of the institution, the collection or the unit, are not related to an annotation type but are also instances of the subclass "Record Concept". These concepts are not annotatable, but are important in their function as identifiers (e.g. to query for records related to a given triple id – the identifier originally used in schemas describing specimens, composed of three ids designating the holding institution, a collection within the institution, and the catalogue number within that collection).
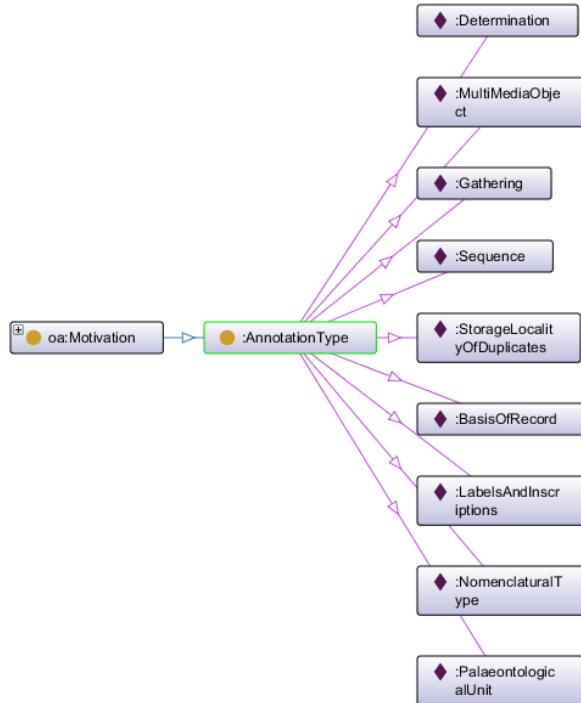
Fig. 2. "Annotation type" is a subclass of *oa:Motivation*, which is a subclass of *skos:concept*.

Table 1. Example concepts of the AnnoSys Ontology and their mappings for annotation type "Determination"

Concept in AnnoSys ontology	Skos exact match	Skos close match	Narrow match	Skos related
Full scientific name	Dwc:Scientific Name	abcd2:TaxonIdentified-FullScientific-NameString		Annotation type: Determination
Scientific Name Authorship	dwc:Scientific NameAuthor-ship		abcd2:TaxonIdentified-AuthorTeam abcd2:TaxonIdentified-AuthorTeamAndYear abcd2:TaxonIdentified-AuthorTeamOriginalAndYear	Annotation type: Determination
Scientific Name Authorship Parenthetical			abcd2:TaxonIdentified-ParentheticalAuthor-TeamAndYear abcd2:TaxonIdentified-AuthorTeamParenthesis abcd2:TaxonIdentified-AuthorTeamParenthesisAndYear	Annotation type: Determination

A prototype of the system is available under https://dev-annosys.bgbm.fu-berlin.de/AnnoSys/AnnoSys.

4 Evaluation

The ontology is composed of around 150 data properties that are related to nine annotation types. Since concepts are now defined in a semantic-web compliant way, they can be stored together with the record in the same triple store (whereas in AnnoSys 1, records have been stored in an XML database). This allows more complex searches and significantly improves the performance of the system. AnnoSys data properties cover the classic annotation workflows in the biodiversity collection data domain. However, the ontology is potentially expandable for other workflows and other domains.

When aiming to integrate annotations for specimens from different data portals, it is essential to be able to identify annotated specimens universally. Therefore, AnnoSys 2 builds persistent identifiers for all objects (records, specimens and annotations) from UUIDs, making the system independent of the previously used tripleIds.

5   Conclusion

Our work tackles the development of an extensible and format-independent system for virtual annotation of biological specimen label data. To this end, we compiled an "AnnoSys-Ontology" mapping essential concepts defined by the widely accepted community standards DarwinCore and ABCD. Annotations are entered via an open browser interface and stored centrally in an RDF triple store following the W3C Web Annotation Data Model.

The system is currently in the testing phase and will be released in 2018. In future research, we will examine the use of AnnoSys for taxon-level data as well as its integration with image annotation systems.

References

1. Duckworth, W.D., Genoways, H.H., Rose, C.L. et al.: Preserving Natural Science Collections: Chronicle of Our Environmental Heritage. National Institute for the Conservation of Cultural Property, Washington, DC. (1993)
2. AnnoSys portal, https://annosys.bgbm.fu-berlin.de/AnnoSys/AnnoSys, last accessed 2017/07/18
3. Berendsohn W.G. (ed.). Access to biological collection data. ABCD Schema 2.06 – ratified TDWG Standard. Berlin: Botanischer Garten und Botanisches Museum Berlin-Dahlem (BGBM), Freie Universität Berlin. (2007)
   http://www.bgbm.org/TDWG/CODATA/Schema/default.htm.
4. DwC terms homepage, http://rs.tdwg.org/dwc/terms/index.htm, last accessed 2017/07/18
5. Tschöpe, O., Macklin, J.A., Morris, R.A. et al. Annotating biodiversity data via the Internet. Taxon, 62, 1248–1258 (2013)
6. Web Annotation Data Model homepage, https://www.w3.org/TR/annotation-model/, last accessed 2017/07/18
7. Suhrbier, L., Kusber, W.-H., Tschöpe, O., Güntsch, A. & Berendsohn, W. G.: AnnoSys - implementation of a generic annotation system for schema-based data using the example of biodiversity collection data. Database (2017). doi:10.1093/database/bax018
8. ABCD2 homepage, https://terms.tdwg.org/wiki/ABCD_2, last accessed 2017/09/07
9. Protege homepage, http://protege.stanford.edu/products.php, last accessed 2017/07/18

A Model to Represent Nomenclatural and Taxonomic Information as Linked Data.
Application to the French Taxonomic Register, TAXREF

Franck Michel[1][0000-0001-9064-0463], Olivier Gargominy[2], Sandrine Tercerie[2] and
Catherine Faron-Zucker[1][0000-0001-5959-5561]

[1] Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France
[2] Muséum national d'Histoire naturelle, Paris, France

Abstract. Taxonomic registers are key tools to help us comprehend the diversity of nature. Publishing such registers in the Web of Data, following the standards and best practices of Linked Open Data (LOD), is a way of integrating multiple data sources into a world-scale, biological knowledge base. In this paper, we present an on-going work aimed at the publication of TAXREF, the French national taxonomic register, on the Web of Data. Far beyond the mere translation of the TAXREF database into LOD standards, we show that the key point of this endeavor is the design of a model capable of capturing the two coexisting yet distinct realities underlying taxonomic registers, namely the nomenclature (the rules for naming biological entities) and the taxonomy (the description and characterization of these biological entities). We first analyze different modelling choices made to represent some international taxonomic registers as LOD, and we underline the issues that arise from these differences. Then, we propose a model aimed to tackle these issues. This model separates nomenclature from taxonomy, it is flexible enough to accommodate the ever-changing scientific consensus on taxonomy, and it adheres to the philosophy underpinning the Semantic Web standards. Finally, using the example of TAXREF, we show that the model enables interlinking with third-party LOD data sets, may they represent nomenclatural or taxonomic information.

Keywords: Linked Data, Taxonomy, Nomenclature, Data Integration.

1 Introduction

Started in the early 2000's, the Web of Data has now become a reality [6]. It keeps on growing through the relentless publication and interlinking of data sets spanning various domains of knowledge. Building upon the Linked Data paradigm [5,14] to connect related pieces of data, this new layer of the Web enables the integration of distributed and heterogeneous data sets, spawning an unprecedented, distributed knowledge graph.

A wealth of existing data sources exists out there, that would valuably populate the Web of Data. For instance, taxonomic registers are key tools to comprehend the diversity of nature and develop natural heritage conservation strategies, *e.g.* by crossing the

myriad records of occurrence data and biological traits. Taxonomic registers are commonly used as the backbone of thematic databases and applications, such as the Global Biodiversity Information Facility[1] that aggregates 54 taxonomic data sources. They may adopt a certain perspective and purpose. For instance, Agrovoc [8] is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization. In this respect, it lists the names of species related to agriculture, fishery and forestry. The NCBI Organismal Classification [12] is another vocabulary covering the organisms specifically referenced in the NCBI nucleotide and protein sequences database. Hence, there does not exist one central register of the taxonomic knowledge. Instead, multiple taxonomic registers cover complementary and often overlapping regions, epochs or domains. Consequently, publishing them as RDF data sets while drawing links between related resources is a way of integrating multiple data sources into a world-scale, biological knowledge graph.

Two coexisting yet distinct realities underlie taxonomic registers, namely the taxonomy (the description and characterization of biological entities called biological taxa, taxon concepts or simply taxa), and the nomenclature (the rules defining how to assign scientific names, or nominal taxa, to these biological entities). The nomenclatural rules are compiled in several Codes. In particular, the Codes for animals [15], plants and fungi [19] and bacteria [17] are used in the TAXREF taxonomic register. The nomenclature yields a controlled thesaurus of scientific names. Each of these scientific names consists of a Latinized name, an authority and a taxonomic rank, along with the original publication and the type specimen bearing that name. Taxonomic registers distinguish each biological taxon from all nominal taxa by retaining a unique reference name for it. For example, taxonomists decided that "*Delphinus capensis* Gray, 1828" and "*Delphinus delphis* Linnaeus, 1758" are the same biological entity, based on morphological or molecular data [10]. In addition to this, the Code of zoological nomenclature rules that this species must be called "*Delphinus delphis* Linnaeus, 1758" as per the principle of priority.

In this paper, we present an on-going work related to TAXREF [13], the French national taxonomic register for fauna, flora and fungus. Our goal is to publish TAXREF on the Web of Data while adhering to standards and best practices for the publication of Linked Open Data (LOD) [11]. First, we analyze how some international taxonomic registers have been published as Linked Data so far. We describe the different modelling choices made to represent the information using the Semantic Web technologies, and the issues that stem from these choices. Then, far beyond the mere translation of the TAXREF database into LOD standards, we show that the key point of this endeavor is the design of a model capable of capturing nomenclatural and taxonomic information. The model we propose has several key advantages: (i) it separates nomenclatural from taxonomic information; (ii) it is flexible enough to accommodate the ever-changing scientific consensus on taxonomy; (iii) it adheres to the philosophy underpinning the Semantic Web standards and it enables drawing links with third-party data sets published as Linked Data, may they represent nomenclatural or taxonomic information.

---

[1] Global Biodiversity Information Facility: https://www.gbif.org/

The rest of this paper is organized as follows. Section 2 analyzes the Linked Data modelling choices of several taxonomic registers. Section 3 describes the model we propose to distinguish between nomenclature and taxonomy. In section 4, we report on more technical aspects of this work, notably the publication of TAXREF according to this model and the production of rich metadata in line with LOD guidelines. Finally, section 5 draws a few conclusions and envisions future actions to be conducted with the biodiversity community.

2   Representing Taxonomic Registers as Linked Data

Several international taxonomic registers have already been published as Linked Data. They adopt somewhat different approaches to model nomenclatural and/or taxonomic information using the Semantic Web stack of technologies. To figure this out, we looked into the following ones: NCBI Organismal Classification [12], Vertebrate Taxonomy Ontology (VTO) [21], Agrovoc Multilingual agricultural thesaurus [8], Encyclopedia of Life (EOL) [7], GeoSpecies Knowledge Base[2] and TaxonConcept Knowledge Base[3]. We also considered the models of two well-adopted generic data sets: DBpedia [18] and BBC Wildlife Ontology[4]. Fig. 1 illustrates the different modelling choices taking the example of the *Delphinus delphis* species and the *Delphinus* genus. Properties with no namespace (*rank* and *genus*) are generic names conveying the idea of such properties; they may be implemented using properties from different ontologies.

- A first option, adopted by NCBI and VTO, is to represent a taxon as an RDFS or OWL class[5] (Fig. 1(a)). The taxonomic ranks are represented by separate classes (*Genus* and *Species* in this example), and a taxon is related to its rank with an appropriate *rank* property. The relationship between a taxon and its parent taxon is modelled by the *rdfs:subClassOf* property.
- Closer to the nomenclature mindset, the model in Fig. 1(b), adopted by Agrovoc, utilizes the SKOS vocabulary[6] to build a thesaurus. Yet, although it could seem that each SKOS concept (an instance of the *skos:Concept* class) solely depicts a scientific name, the model embeds synonymy relationships that are typical of taxonomic information. The child-to-parent relationship between two scientific names is represented by the *skos:broader* property.
- The EOL database is queried by means of an API[7] that returns results in the RDF JSON-LD syntax. A response makes use of the Darwin Core standard for biodiversity data exchange [23]: each taxon is rendered as an instance of the *dwc:Taxon* class, as depicted in Fig. 1(c), that is meant to denote taxonomic information (*dwc:Taxon*

---

[2]   https://bioportal.bioontology.org/ontologies/GEOSPECIES
[3]   http://lod.taxonconcept.org/
[4]   BBC Wild Life Ontology : http://www.bbc.co.uk/ontologies/wo
[5]   OWL2: https://www.w3.org/TR/2012/REC-owl2-rdf-based-semantics-20121211/
[6]   SKOS: https://www.w3.org/2009/08/skos-reference/skos.html
[7]   EOL API : http://eol.org/info/api_overview

is equivalent to *Taxon* and *TaxonConcept* in the TDWG Ontology[8]). Nomenclatural information is hardly separated from taxa.

— The model in Fig. 1(d) defines specific classes for each taxonomic rank, such as *Species* and *Genus*. Unlike models (a) to (c), the taxonomic rank is not denoted by a specific property but by the belonging to a class, *e.g. Delphinus delphis* is an instance of the *Species* class. The child-to-parent relationship is represented by a per-rank property, *genus* in this case. This model has been adopted by GeoSpecies, DBpedia and the BBC Wildlife Ontology.

— Lastly, TaxonConcept's model (Fig. 1(e)) is very similar to model (d), with the difference that only the species rank is represented as a class. Higher ranks are simply mentioned by means of a per-rank property whose object is a literal (property *genus* and literal "*Delphinus*" in the example).
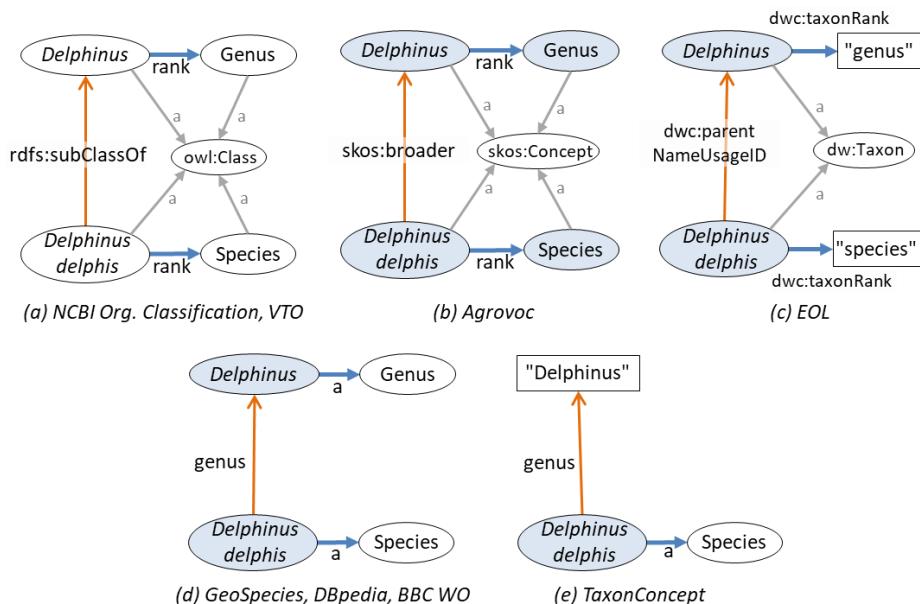
Fig. 1. Various models to represent taxa and/or scientific names using OWL classes (a), SKOS concepts (b) or instances of other classes (c, d and e). Boxes depict literals. White bubbles are OWL classes whereas blue bubbles are class instances. Orange arrows depict the child-to-parent relationship between the *Delphinus delphis* species and the *Delphinus* genus. Blue arrows relate a taxon with a taxonomic rank.

In spite of these differences, all those models seem to depict the same reality. Nevertheless, a careful look suggests that they convey somewhat varying mindsets. In the Semantic Web ethos, OWL classes are defined by extension as a set of instances (or individuals). Intuitively, the *Delphinus delphis* class in (a) comprises the individuals of

---

8   https://github.com/darwin-sw/dsw/wiki/ClassTaxon#equivalence-of-taxon-and-taxoncon-cept-in-the-tdwg-ontology-and-the-darwin-core-standard

that species. This is in line with the models of NCBI and VTO that mostly provide a biological description, *i.e.* taxonomic information. By contrast, SKOS is commonly used to describe a nomenclatural system as a thesaurus, *i.e.* a hierarchy of concepts connected by semantic relationships. Yet, the generic term "nomenclatural system" must not be confused with the nomenclature in its biological sense. Indeed, Agrovoc (b) models a hierarchy of concepts that not only represent nomenclatural information (scientific names) but also taxonomic information (how names are assigned to taxa) intertwined with each other. Similarly, EOL (c) chooses to model a taxon as an instance of the *dwc:Taxon* class. Using OWL classes on the one hand, or SKOS concepts or *dwc:Taxon* instances on the other hand, are equally valid solutions. Only, they indicate different perspectives of the same reality: an instance (in particular of the *skos:Concept* class) characterizes a taxon as one concept within a thesaurus of taxon concepts, while an OWL class characterizes a taxon as the set of individuals of that biological entity. The use of instances to represent species in GeoSpecies (d) and TaxonConcept (e) makes them close to the SKOS mindset. Both describe scientific names along with occurrence data, thus, again, interweaving taxonomic and nomenclatural information.

Hence, to some varying extent, it occurs that all these approaches intertwine taxonomic information and nomenclatural information. When we consider a broader picture, these discrepancies entail several impediments:

— Firstly, the scientific consensus about taxonomy constantly evolves. For instance, Linné described most snails as species belonging to genus *Helix* in 1758, but many of them now belong to another family, *e.g.* "*Helix glauca* Linnaeus, 1758" is a synonym of "*Pomacea glauca* (Linnaeus, 1758)" which is the valid name. Similarly, "*Delphinus capensis* Gray, 1828" became a synonym of "*Delphinus delphis* Linnaeus, 1758" in 2015 in light of new scientific evidences [10]. When nomenclatural and taxonomic information is intertwined, the model pictures a snapshot of the use of scientific names at a certain time, that can hardly accommodate changes. A workaround to this issue consists in versioning the whole data set but this entails setting up a mechanism to track the changes from one version of the data set to the next. Editorial notes can be used to document such changes but these are mainly meant for humans and are hardly machine-processable. For a model to accommodate such changes in a flexible manner, it is necessary to distinguish explicitly between the nomenclatural and taxonomic levels. This distinction may allow not only to follow up on taxonomical changes, but also to track and characterize them as proposed by Chawuthai et al [9].

— Secondly, the power of Linked Data spawns from the number and quality of links. Interlinking two data sets requires that they model the same kind of information. If it is unclear whether the focus of a data set is about nomenclature (scientific names) or biology (taxa), then drawing *owl:sameAs* links with resources of other data sets may be erroneous: a species name is not the same thing as the group of individuals of that species. Furthermore, a more technical limitation can occur when interlinking data sets: good practices generally discourage the alignment of class instances with classes since reasoners for Description Logics rely on the distinction between terminological and assertional knowledge [1]. Interestingly enough, this issue is strikingly

evidenced by the data sets that we analyzed: NCBI and VTO, both based on OWL classes, are linked with each other using the *owl:equivalentClass* property, but they have no link whatsoever with the data sets based on instances[9] (models b, c, d and e of Fig. 1). This absence of links does not result from a conceptual mismatch; it results from a sheer technical issue, although conceptually, it would make perfect sense to link NCBI and VTO with these other data sets.

In the next section, we propose a model intended to tackle these issues in the context of the TAXREF taxonomic register.

3 A Generic Model to represent Nomenclatural and Taxonomic Information as Linked Data

TAXREF [13] is the French national taxonomic register for fauna, flora and fungus, maintained and distributed by the National Museum of Natural History of Paris (France). It is a manually curated register of all the species inventoried in metropolitan France and overseas territories, organized as a hierarchy of over 500.000 scientific names that mark a national and international consensus. From the temporal perspective, all living beings are considered as well as those of the close natural history, from the Paleolithic until now. Available through a Web site[10], a Web service[11] or a down-loadable text file, TAXREF enables the interoperability between biological databases (mainly occurrence databases), thus supporting biodiversity studies and natural heritage conservation strategies. A new version of TAXREF is published every year, that acknowledges synonymy or hierarchy changes.

Our goal is to design a model to represent TAXREF as Linked Data, that works out the issues and limitations discussed in section 2. More specifically, we seek to achieve three objectives:

1. the model must be relevant to biologists by reflecting the distinction between nomenclature and taxonomy, as well as to computer scientists by adhering to the philosophy that underpins the Semantic Web standards;
2. the model must be flexible enough to accommodate taxonomic changes from one version of TAXREF to the next;
3. the model must enable the alignment with third-party data sets published as Linked Data, may they represent nomenclatural or taxonomic information.

Fig. 2 sketches the model we propose to publish TAXREF as Linked Data, that we denote TAXREF-LD. It is the outcome of a thorough reflection during which we confronted concepts from the biology (taxonomy, systematics) with Semantic Web modelling practices and LOD publication pragmatic concerns.

---

[9] Here we refer to proper LOD links using HTTP URIs. NCBI and VTO embed cross references to third-party database identifiers (using *e.g.* property *obo:hasDbXref*), but these do not comply with LOD principles.

[10] https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=en

[11] https://taxref.mnhn.fr/taxref-ws
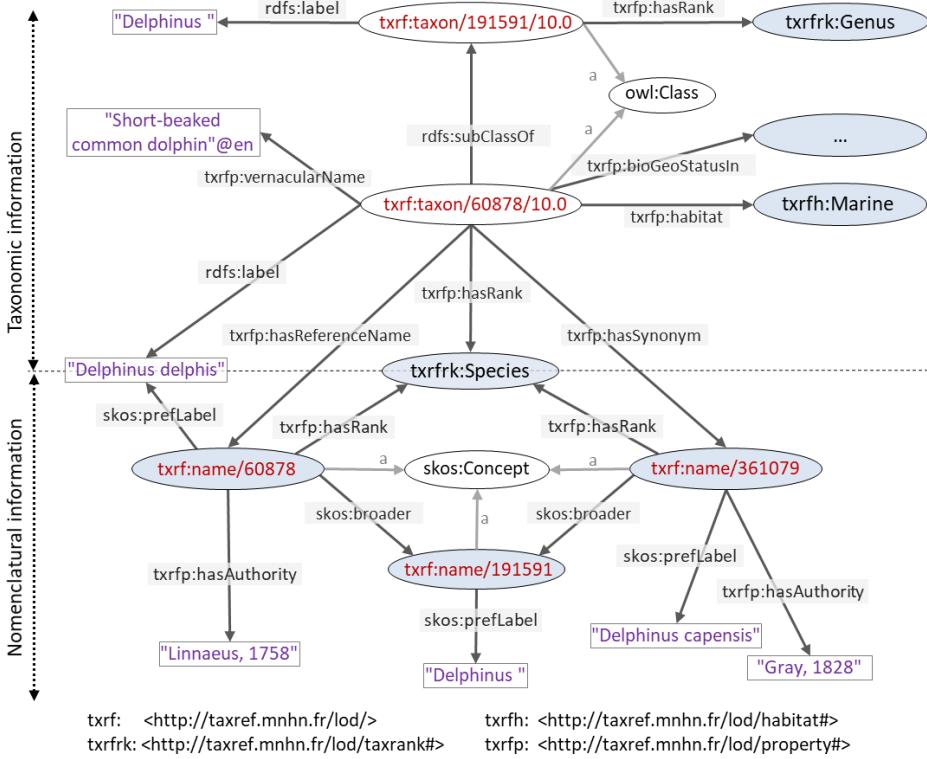
Fig. 2. TAXREF-LD: a Linked Data model for TAXREF, separating nomenclatural from taxonomic information. White bubbles are OWL classes whereas blue bubbles are class instances.

To achieve objective 1, two distinct levels are modelled. At the nomenclatural level (lower part of Fig. 2), each scientific name is modelled as a SKOS concept along with a label, an authority and a taxonomic rank. The child-to-parent relationship between scientific names is expressed with the *skos:broader* property. At the taxonomic level (upper part), a biological taxon is modelled as an OWL class. As a mirror of the nomenclatural level, the child-to-parent relationship between taxa is expressed with the *rdfs:subClassOf* property. Vernacular names are not governed by nomenclatural rules, but account for a property of a group of individuals. Hence, they are attached to the taxon's OWL class. Likewise, biological traits (currently habitat and biogeographical status) are attached to the OWL class. Both levels are connected by the links between a taxon and its reference name (property *txrfp:hasReferenceName*), and between the taxon and its synonyms (property *txrfp:hasSynonym*). A taxon gets the label of its reference name, hence the *rdfs:label* property with the same value as the reference scientific name's *skos:prefLabel* property. It also takes the taxonomic rank of its reference name, hence the *txrfp:hasRank* property at both levels.

In this model, instances of a taxon's OWL class are not depicted. They would typically be the biological individuals of that taxon. In particular, an instance may be created to represent and characterize the type specimen attached to a scientific name.

Note that, for the sake of clarity, details of biogeographical statuses are not depicted in Fig. 2. Also, taxonomic ranks and types of habitats are instances of the *skos:Concept* class but this is not depicted.

OWL class vs. Darwin Core Taxon. Arguably, an alternative model could represent taxa as instances of the *dwc:Taxon* class, rather than OWL classes. The Darwin Core terms were initially designed as a means to exchange taxonomic data using flat text files. As of today, the journey towards a proper ontological representation in RDF is still on-going, as pointed out by Baskauf et al [4]. Despite efforts of the Darwin-SW project to define object properties relating organisms, identifications, taxa, occurrences and locations [2], some issues have not been addressed yet, as underlined in [3]: "the object properties necessary to relate *dwc:Taxon* instances to name entities, references, parent taxa, and child taxa do not exist and the exact relationship between taxonomic entities such as taxon concepts, protonyms, taxon name uses, etc. has not been established using RDF". Accordingly, it occurred to us that the RDF representation of Darwin Core terms is not mature enough yet to fulfill the distinction we wish to model between the nomenclatural and taxonomic information levels.

URI naming scheme. The nomenclatural level is stable in time: new scientific names may be coined but the information associated with a name shall not change, as ruled by the Codes of nomenclature. Consequently, URIs of SKOS concepts are fixed once for all versions of TAXREF. For instance, *Delphinus capensis* is associated a SKOS concept whose URI is *http://taxref.mnhn.fr/lod/361079/name*. Conversely, the taxonomic level must be able to accommodate changes (objective 2). Our point is not to characterize and keep track of the changes that may occur through time (in contrast to *e.g.* [9]), but simply to allow changes in the use of scientific names by taxon concepts, between two versions of TAXREF. Toward this end, we append TAXREF's version number to the URIs of OWL classes. For instance, *Delphinus capensis* was a reference name in version 9.0, thus it was associated an OWL class (*http://taxref.mnhn.fr/lod/taxon/361079/9.0)* and a SKOS concept (given above). Since version 10.0, it has become a synonym of *Delphinus delphis*, hence it has no corresponding OWL class in version 10.0, only the SKOS concept remains.

Interlinking. The separate modelling of the nomenclatural and taxonomic levels provides greater flexibility for the interlinking with third-party data sets (objective 3). For instance, NCBI's classes model biological taxa that are linked with TAXREF-LD's taxonomic level using the *owl:equivalentClass* property (section 4 discusses further the choice of relevant linking properties). The distinction between nomenclatural and taxonomic levels may also be useful to avoid linking biological entities that bear the same scientific name although they denote different entities throughout data sets. For example, the IUCN Red List of Endangered Species[12] still considers *Delphinus delphis* and *Delphinus capensis* as separate species, although *Delphinus capensis* is now considered as a synonym for *Delphinus delphis*. Consequently, 'their' *Delphinus delphis* taxon does not denote the same biological entity as the one in TAXREF, thence a link at the taxonomical level would be erroneous. Yet, a link at the nomenclatural level (names) makes sense since it does not depend on synonymy relationships.

---

[12] http://www.iucnredlist.org

4   Publishing TAXREF-LD as High Quality Linked Data

To perform the translation of the TAXREF database into the model presented in section 3, we used the Morph-xR2RML software[13], an implementation of the xR2RML generic mapping language [20] designed to address the translation of heterogeneous data sources into RDF. This produced a graph of approximately 8.5M RDF triples, accounting for 509.148 scientific names (SKOS concepts) and 236.507 taxa (OWL classes).

Access methods. An on-going work intends to set up a server enabling the sustainable dereferencing of TAXREF-LD URIs. Until then, a temporary server hosts the RDF graph for test purposes. It provides a dereferencing method[14] as well as a public SPARQL endpoint[15].

Metadata. In order to ensure discoverability, understandability and exploitability of TAXREF-LD, we have taken great care of providing rich and informative metadata while adhering to best practices for the publication of data on the Web [11]. Using the DCAT vocabulary[16], we defined a catalog (*http://taxref.mnhn.fr/lod/TaxrefCatalog*) wherein the different versions of TAXREF are represented by separate DCAT data sets. Each data set comes with three distributions: a Web service, a downloadable text file and a Linked Data distribution *i.e.* TAXREF-LD (*http://taxref.mnhn.fr/lod/Taxref-ld/10.0* in TAXREF version 10.0). Additional annotations are provided with respect to the number of triples, vocabularies used, links with other data sets, provenance, etc., using notably the VoID vocabulary[17]. The TAXREF-LD resource is also the SKOS thesaurus (of type *skos:ConceptScheme*) that registers all the SKOS concepts representing scientific names. *Biota* (*http://taxref.mnhn.fr/lod/name/349525*) is its top concept.

Links with other taxonomic registers. To achieve significant interlinking, we first manually aligned the TAXREF-LD classes and properties (related to taxonomical ranks, habitats, authority, etc.) with their counterparts from other ontologies. Then, we developed a plugin for the Silk Framework [22], that ports a matching algorithm previously developed by TAXREF experts. We leveraged the distinction between the nomenclatural and taxonomic levels to link TAXREF-LD with datasets based on the multiple models presented in Fig. 2. NCBI Organismal Classification and VTO both define classes that we aligned with the taxonomic level of TAXREF-LD, as illustrated in the upper part of Fig. 3. With a model based on SKOS concepts, Agrovoc's SKOS concepts are more likely linked with TAXREF-LD's nomenclatural level using the *skos:exactMatch*. Yet, this equivalence is controversial since taxonomic information is interweaved in Agrovoc's model. An alternative may be to use the weaker *skos:closeMatch* property, or to assume that Agrovoc's concepts represent taxa and declare TAXREF-LD's SKOS concepts as reference or synonymous names of these taxa. Likewise, with

---

[13]   Morph-xR2RML: https://github.com/frmichel/morph-xr2rml/
[14]   Any TAXREF-LD URI can be dereferenced by pointing to http://erebe-vm2.i3s.unice.fr:8890/describe/?url=<URI>. For instance, this tiny URL leads to the description of taxon *Delphinus delphis*: https://frama.link/RJd-_xq8
[15]   http://erebe-vm2.i3s.unice.fr:8890/sparql
[16]   DCAT: https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
[17]   VoID: https://www.w3.org/TR/2011/NOTE-void-20110303/

an instance-based modelling of taxa intertwined with some nomenclatural information, TaxonConcept and GeoSpecies are controversial cases. As discussed in section 2, good practices recommend not to align these instances with OWL classes of TAXREF-LD's taxonomic level, unless utilizing a semantically-poor property such as *rdfs:seeAlso*. Thus, we opted for an alignment at the nomenclatural level of TAXREF-LD, yet using the weaker *skos:relatedMatch* or *skos:closeMatch* SKOS properties, depending on how close they are to our model. This is illustrated in the lower part of Fig. 3. The linking with EOL is still on-going at the time of writing. Overall, we created 267.155 links towards resources of these taxonomic registers. Additionally, TAXREF maintains references to Web pages of on-line scientific databases. We used these references to produce 992.722 *foaf:page* links from TAXREF-LD classes and concepts towards related Web pages (not depicted in Fig. 3).
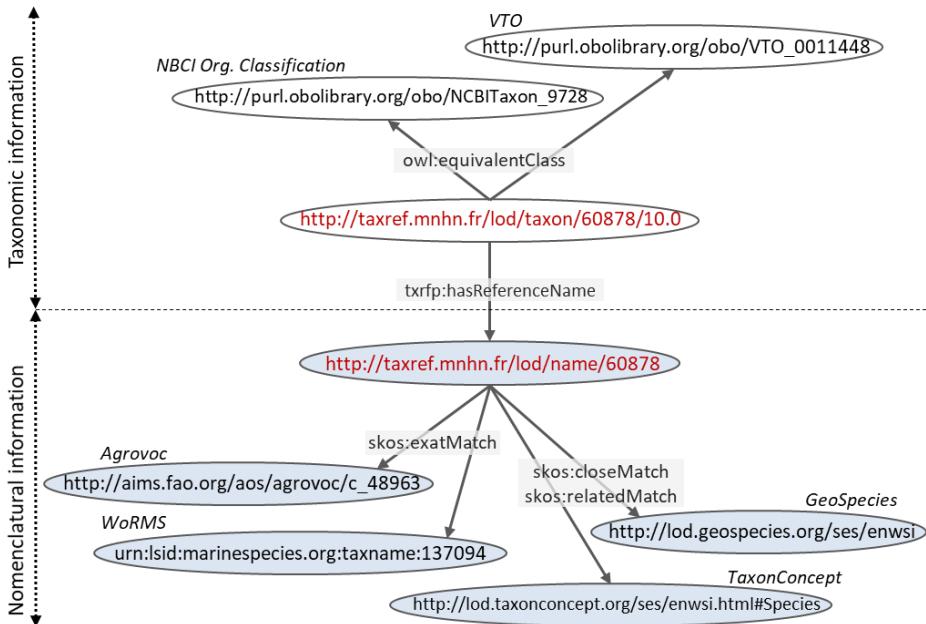
Fig. 3. Interlinking of the *Delphinus delphis* species with six other LD taxonomic registers

5   Conclusion and Perspectives

Taxonomic registers are key tools for the integration of biological databases. As such, they stand out as promising candidates to populate the Web of Data. In this paper, we reported on the publication of the French taxonomic register (TAXREF) in the Web of Data, by adhering to Linked Open Data best practices.

We first analyzed the varying modelling choices made in the past years to represent some international taxonomic registers as Linked Data. We pointed out that these models convey different mindsets that can make their interlinking difficult. Furthermore,

these models do not easily accommodate the ever-changing scientific consensus about taxonomy.

Consequently, we proposed a model tackling these issues and capable of capturing two distinct levels of information: nomenclatural information (scientific names assigned to biological entities) is represented as a SKOS thesaurus, and taxonomic information (the description and characterization of these biological entities) is represented by OWL classes. We argue that this model is relevant to biologists as well as Semantic Web experts, it is flexible enough to accommodate taxonomy changes and it enables interlinking with third-party data sets published as Linked Data, whatever the model they adopted. We applied this model to the case of TAXREF, that is now publicly accessible through a SPARQL endpoint and a Linked Data server, and we seek to achieve proper dereferencing of the URIs in the near future. To increase its visibility, we are in the process of registering TAXREF-LD on the DataHub.io portal, and we are considering its publication on the AgroPortal ontology repository for agronomy [16] .

Furthermore, our goal with this paper is to engage in a discussion with the stakeholders of the biodiversity community, may they be data consumers or producers of sibling taxonomic registers covering complementary regions, epochs or domains. Our point it to delineate some scientific questions and the underlying data integration scenarios, and engage in actions to pursue these objectives.

More generally, the publication of taxonomic registers as Linked Data is a way to contribute to a large, distributed, biological knowledge base. This knowledge base may be beneficial in many ways. For instance, taxonomists may leverage it to compare and discuss their conceptions of biological entities throughout the world. Navigating through interlinked data sets related to domains as diverse as the biology, genetics, medicine, resources management, sociology, etc., could pave the way to inferring new knowledge on organisms and spur new research areas.

References

1. F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider: The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press, New York, NY, USA (2003).
2. S.J. Baskauf, C.O. Webb: Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF, Semantic Web. 7 (2016) 617–627.
3. S.J. Baskauf, J. Wieczorek, J. Deck, C. Webb, P.J. Morris, M. Schildhauer: Darwin Core RDF Guide, Biodiversity Information Standards. (2015).
4. S.J. Baskauf, J. Wieczorek, J. Deck, C.O. Webb: Lessons learned from adapting the Darwin Core vocabulary standard for use in RDF, Semantic Web. 7 (2016) 617–627.
5. T. Berners-Lee: Linked Data, in Design Issues of the WWW, (2006).
6. C. Bizer: The Emerging Web of Linked Data, IEEE Intelligent Systems. 24 (2009) 87–92.
7. R. Blaustein: The Encyclopedia of Life: Describing Species, Unifying Biology, BioScience. 59 (2009) 551–556.

8. C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, et al.: The AGROVOC linked dataset, Semantic Web. 4 (2013) 341–348.

9. R. Chawuthai, H. Takeda, V. Wuwongse, U. Jinbo: Presenting and Preserving the Change in Taxonomic Knowledge for Linked Data, Semantic Web. 7 (2016) 589–616.

10. H.A. Cunha, R.L. de Castro, E.R. Secchi, E.A. Crespo, J. Lailson-Brito, A.F. Azevedo, et al.: Molecular and Morphological Differentiation of Common Dolphins (Delphinus sp.) in the Southwestern Atlantic: Testing the Two Species Hypothesis in Sympatry, PLOS ONE. (2015).

11. B. Farias Lóscio, C. Burle, N. Calegari: Data on the Web Best Practices, W3C Recommandation. (2017).

12. S. Federhen: The NCBI Taxonomy database, Nucleic Acids Research. 40 (2012) D136–D143.

13. O. Gargominy, S. Tercerie, C. Régnier, T. Ramage, C. Schoelink, P. Dupont, et al.: TAXREF v10. 0, référentiel taxonomique pour la France: méthodologie, mise en oeuvre et diffusion, Muséum National d'Histoire Naturelle, Paris. (2016).

14. T. Heath, C. Bizer: Linked Data: Evolving the Web into a Global Data Space, 1st ed., Morgan & Claypool, (2011).

15. International Commission on Zoological Nomenclature: International Code of Zoological Nomenclature, Fourth Edition, International Trust for Zoological Nomenclature, (1999).

16. C. Jonquet, A. Toulet, E. Arnaud, S. Aubin, E. Dzalé-Yeumo, V. Emonet, et al.: Reusing the NCBO BioPortal technology for agronomy to build AgroPortal. In Proceedings of the 7th International Conference on Biomedical Ontologies, ICBO'16, Demo Session, Corvallis, Oregon, USA (2016).

17. S.P. Lapage, P.H. Sneath, E.F. Lessel, V.B.D. Skerman, W.A. Clark, H.P.R. Seeliger: International Code of Nomenclature of Bacteria: Bacteriological Code - 1990 Revision, ASM Press, (1992).

18. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, et al.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, Semantic Web. 6 (2014) 167–195.

19. J. McNeill, F.R. Barrie, W.R. Buck, V. Demoulin, W. Greuter, D.L. Hawksworth, et al.: International Code of Nomenclature for algae, fungi, and plants (Melbourne Code). Regnum Vegetabile 154, Koeltz Scientific Books, (2012).

20. F. Michel, C. Faron-Zucker, J. Montagnat: Translation of Heterogeneous Databases into RDF, and Application to the Construction of a SKOS Taxonomical Reference. In Revised Selected Papers of the 11th International Conference on Web Information Systems and Technologies (WebIST), Springer, (2016): pp. 275–296.

21. P.E. Midford, T.A. Dececchi, J.P. Balhoff, W.M. Dahdul, N. Ibrahim, H. Lapp, et al.: The Vertebrate Taxonomy Ontology: a framework for reasoning across model organism and species phenotypes, Journal of Biomedical Semantics. 4 (2013) 34.

22. J. Volz, C. Bizer, M. Gaedke, G. Kobilarov: Silk - A Link Discovery Framework for the Web of Data. In 2nd Workshop about Linked Data on the Web, Madrid, Spain (2009).

23. J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, et al.: Darwin Core: An Evolving Community-Developed Biodiversity Data Standard, PLOS ONE. 7 (2012).

Bottom-up taxon characterisations with shared knowledge: describing specimens in a semantic context

Patrick Plitzner[1][0000-0002-7740-5423], Tilo Henning[1], Andreas Müller[1], Anton Güntsch[1], Naouel Karam[2] and Norbert Kilian[1]

[1] Botanic Garden and Botanical Museum Berlin, Freie Universität Berlin, Germany
[2] Department of Mathematics and Computer Science, Freie Universität Berlin, Germany
p.plitzner@bgbm.org

Abstract. Using the angiosperm order Caryophyllales, we will provide an exemplar use case on optimizing the taxonomic research process with respect to delimitation and characterisation ("description") of taxa using the the European Distributed Institute of Taxonomy (EDIT) Platform for Cybertaxonomy. The workflow for sample data handling of the EDIT platform will be extended: Character data (data on genotypic and phenotypic characters of any type, here focusing on morphology) will be captured and stored in structured form. The structure consists of character and character state matrices for individual specimens instead of taxa, which shall allow to generate taxon characterisations by aggregating the data sets for the individual specimens included. To ensure data integrity, especially for the aggregation process, semantic web technologies will be used to establish and continuously elaborate expert community-coordinated exemplar vocabularies with term ontologies and explanations for characters and states. In cooperation with the "German Federation for Biological Data" (GFBio), the GFBio Terminology Service is used for publishing the ontologies via a public API. The EDIT platform will be extended to use and integrate the GFBio Terminology Service in order to work with the latest version of the ontology used for specimen respective taxon descriptions.

Keywords: descriptive data, e-taxonomy, terminology management

1 Pre-work and project goals

In a precursor project [1, 2], we have implemented a workflow for processing specimen-related metadata on the **E**uropean **D**istributed **I**nstitute of **T**axonomy (**EDIT**) Platform for Cybertaxonomy [3], a comprehensive taxonomic data management and publication environment that offers a collection of tools and services and works as a service provider to support taxonomic workflows, publishing, data storage and exchange, etc. The aim was to organise the links between (a) samples of individual organisms collected, (b) research data obtained from them, (c) specimens of these individuals deposited in research collections, and (d) taxon assignments ("identifications") of the investigated individuals.

On this basis, the current project will optimise the taxonomic research process with respect to delimitation and characterisation ("description") of taxa.

Working on the angiosperm order Caryophyllales [4], character data (mainly morphological data) of individual specimens will be recorded and stored in the underlying "Common Data Model" (CDM) [5] compliant data store of the platform. For specimen descriptions, a community-developed expert ontology backed by the GFBio terminology service for ontology management is being developed and used to ensure data integrity. In a final step, data aggregation of the individual character data sets assisted by the terminology service will generate automated descriptions on taxon level.

This project combines two major scientific areas, semantic descriptions and taxon characterization both of which are crucial for sustainable scientific work. Taxon characterizations on specimen level allow for generated taxonomic delimitation. However, this is partly a subjective work leading to different definitions for certain features (leaf colour is "reddish green" vs "greenish red"). To align different characterizations the combination with semantically defined terms will relate existing definitions and also unify newly created ones by proposing existing terms.

2   Terminology service

One of the project goals is to create an ontology for specimen descriptions which should be used and developed collaboratively. This ontology should be made publicly available to increase the reach and usage of the semantic concepts developed for it. The GFBio terminology service [6], which is simultaneously being implemented, supports working with formal ontologies, taxonomies or other Semantic Web compliant collections of terms. It will be used to store and publish the aforementioned ontology. The service, as seen in Fig 1, provides a web service interface to support various requests related to retrieving semantic information from the stored ontologies. Another important feature is the mapping of internal and external terminological resources which promotes even more the collaborative work on ontologies.
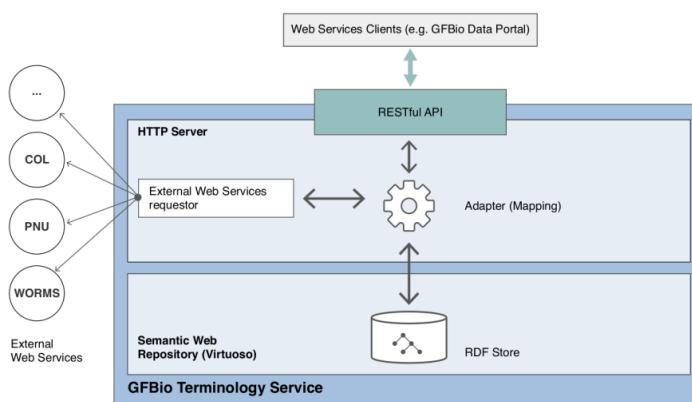
Fig 1 Overview of the GFBio terminology service architecture

3   Specimen Description workflow

Ontologies backed by the terminology service will be created, managed, used and extended during the entire workflow for specimen based data acquisition and taxon descriptions. Three main applications can be identified, all of which will be integrated into the EDIT platform as part of the current project (see Fig 2)
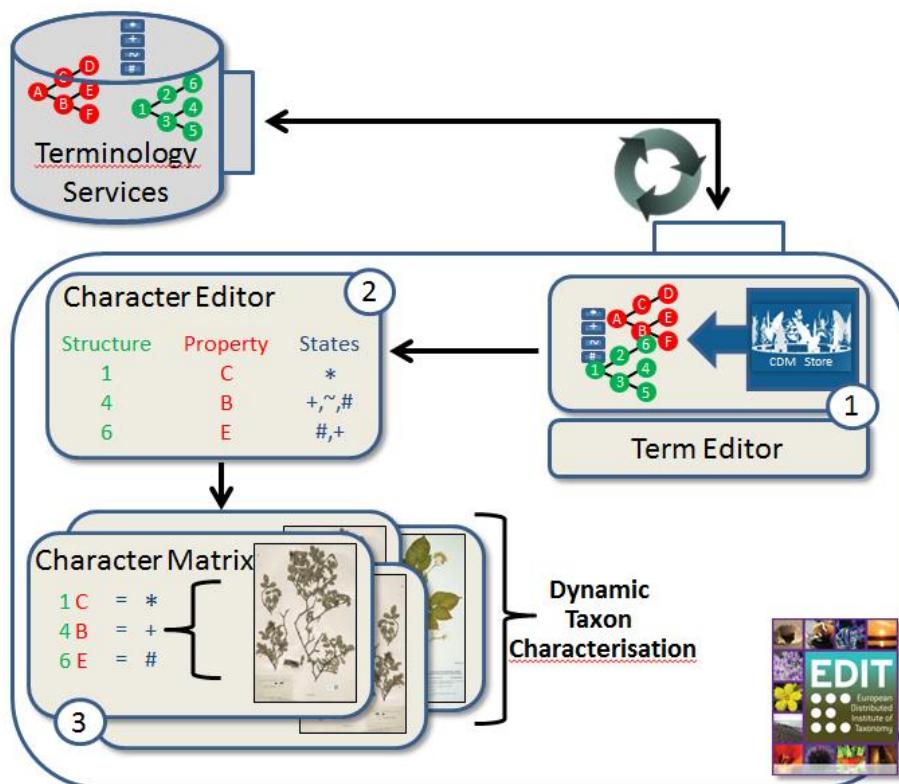
Fig 2 The EDIT platform uses the API of the terminology service to integrate the terminology services into three applications: 1) the term editor which allows editing on a synced copy of the ontology, 2) the character editor where the user defines taxon specific term hierarchies for structures, properties and their corresponding states and 3) the character matrix which serves for the character-based description of single specimens.

3.1   Ontology Management

Ontology editing facilities are implemented into the EDIT platform using the API of the terminology service. The platform itself provides a user and rights management which will serve for collaborative work on the ontology preparation and maintenance.

Additionally, the CDM as the storage model adds more fine-grained meta information to the development process. It allows tracking changes i.e. allowing a versioning mechanism and also an extended documentation via annotations and notes is possible.

Working on the ontology within the platform will be done on a synced copy of the data. The CDM will be extended to support the linkage of terms and their relations as well as their semantic concept in the remote ontology provided by the terminology service.

A term editor based on the EDIT platform is used to visualise and edit the synced copy.

3.2   Creating the descriptive data set/Character editor

For a comprehensive morphological analysis of a taxon in general as well as specimen-wise, a well-defined, established terminology is essential that has already been widely used in the respective plant group. The individual botanist must be able to choose the necessary terms from a vocabulary that is persistently embedded in or linked to a stable term-ontology (e.g. The Plant Ontology [7]).

To describe the morphological characters observed, composite terms are used following the tripartite principle proposed by Diederich [8] and realised in the Prometheus model [9, 10]. That means that characters are composed of three single terms that belong to different categories: (1) plant structures, defining the morphological structure of a plant organism from root to flower, (2) properties, describing the morphological aspects of the plant structures, (3) states for setting the quantitative or categorical space of the properties.

Structures and properties will be stored in *tree structures* into CDM-based data stores. The tree structure allows for designing taxonomic group specific hierarchies and dependencies between the single terms. The compilation of structure tree, property tree and states connected to a taxon is called a *descriptive data set*.

3.3   Character matrix and aggregation

The first two steps dealt with the conceptual creation of the descriptive data set by evaluating what terms of the ontology are needed, how they are ordered and how their boundaries are defined. The final step is the actual description of specimens including the creation of characters and measuring their states.

As pointed out in the previous chapter, data triplets based on the Prometheus model are used. Every single character that describes a certain feature of the specimen is built up from a structure term and a property term. The range of the property term itself is limited by the states assigned to it.

The specimen descriptions are edited in a *character matrix* combining all specimens associated with the taxonomic group of the current descriptive data set with the characters created to describe the morphological features. The matrix can be seen as a table with ordered rows which will be built up by the characters that were previously created to describe the taxon. The columns will be the specimens belonging to that certain

taxon. The order of the characters also provides semantic information. There are, for example, character that cannot exist because the overall structure to which they belong does not exist as well as a more general character may already define the boundaries of a sub character.

The editing process will be enriched with the semantic knowledge about the terms. This enables rules for value hierarchies, data entry assistance through semantic documentation, data validation, etc.

The ordering of state information into a character matrix enables the procedure of generating taxon descriptions via an aggregation algorithm. Specimen descriptions will be comparable to each other because of structured character data organization. Single characters and their states are semantically defined by the underlying ontology describing what they are and how to interpret their values. The semantic knowledge also assists when comparing or merging character data from different sources.

4   Conclusion and Future Work

The EDIT platform in combination with the GFBio terminology service creates a capable environment for the process of a specimen-based and dynamic description of taxa using character data. The descriptive data set as a data structure connects the "raw" specimen character data to a taxonomic group, making data aggregation possible which allows the generation of automated taxon descriptions. Each application of the workflow is based on the platform and the CDM so that the user rights and roles management system can be set up specifically for each task by granting access only to those users that are authorized.

In any step of the workflow it is common that requests to change or edit the ontology will come up. The CDM provides the link to the synced copy of the ontology but anyway, in a future step, change and versioning strategies should be discussed in more detail as there are still no established solutions to this problem.
Another advantage of working with semantic technology is reasoning. This will especially be of interest during the aggregation process when dealing with conflicting data or generated taxon descriptions vs. descriptions from literature.

References

1. Kilian N, Henning T, Plitzner P, Müller A, Güntsch A, Stöver BC, Müller KF, Berendsohn WG, Borsch T (2015) Sample data processing in an additive and reproducible taxonomic workflow by using character data persistently linked to preserved individual specimens. Database 2015: 1–19. doi:10.1093/database/bav094
2. Campanula Data Portal. http://campanula.e-taxonomy.net/
3. Berendsohn WG (2010) Devising the EDIT Platform for Cybertaxonomy. In: Nimis L, Vignes-Lebbe R (eds). Tools for Identifying Biodiversity: Progress and Problems. roceedings of the International Congress, Paris, 20–22 September 2010. EUT Edizioni niversita` di Trieste, Trieste, pp. 1–6.

4. Borsch T, Hernandez-Ledesma P, Berendsohn WG, Flores-Olvera H, Ochoterena H, Zuloaga FO, v. Mering S, Kilian N (2015) An integrative and dynamic approach for monographing species-rich plant groups—building the global synthesis of the angio-sperm order Caryophyllales. Perspect Plant Ecol Evol Syst 17: 84–300. doi.org/10.1016/j.ppees.2015.05.003

5. Anonymous. (2008) Common Data Model. http://dev.e-taxonomy.eu/trac/wiki/CommonDataModel (25 July 2017, date last accessed).

6. Naouel Karam, Claudia Müller-Birn, Maren Gleisberg, David Fichtmüller, Robert Tolksdorf, Anton Güntsch: A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data. Datenbank-Spektrum 16(3): 195-205 (2016)

7. The Plant Ontology. http://planteome.org/

8. Diederich J (1997) Basic properties for biological databases: character development and support. Math Computer Model 25: 109–127.

9. Pullan MR, Watson MF, Kennedy JB, Raguenaud C, Hyam R (2000) The Prometheus Taxonomic Model: A Practical Approach to Representing Multiple Classifications. Taxon 49(1): 55–75.

10. Pullan MR, Armstrong KE, Paterson T, Cannon A, Kennedy JB, Watson MF, McDonald S, Raguenaud C (2005) The Prometheus Description Model: an examination of the taxonomic description-building process and its representation. Taxon 54(3): 751–765.

Adding Biodiversity Datasets from Argentinian Patagonia to the Web of Data

Marcos Zárate[1,2,4] Germán Braun[3,4] Pablo Fillottrani[5,6]

[1] Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico (CESIMAR-CENPAT), Argentina
[2] Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Argentina
[3] Universidad Nacional del Comahue (UNCOMA), Argentina
[4] Consejo Nacional de Invenstigaciones Científicas y Técnicas (CONICET), Argentina
[5] Universidad Nacional del Sur (UNS), Argentina
[6] Comisión de Investigaciones Científicas de la provincia de Buenos Aires (CIC), Argentina

Abstract In this work we present a framework to publish biodiversity data from Argentinian Patagonia as Linked Open Data (LOD). These datasets contains information of biological species (mammals, plants, parasites, among others) have been collected by researchers from the Centro Nacional Patagónico (CENPAT), and have initially been made available as Darwin Core Archive (DwC-A) files. We introduce and detail a transformation process and explain how to access and exploit them, promoting integration with other repositories.

Keywords: Biocollections, Darwin Core, Linked data, RDF, SPARQL

1 Introduction

Animal, plant and marine biodiversity comprise the "natural capital" that keeps our ecosystems functional and economies productive. However, since the world is experiencing a dramatic loss of biodiversity [1,2], an analysis about its impact is being done by digitising and publishing biological collections [3]. To this end, the biodiversity community has standardised shared common vocabularies such as Darwin Core (DwC) [4] together with platforms as the Integrated Publishing Toolkit (IPT) [5] aiming at publishing and sharing biodiversity data. As a consequence, the biodiversity community now have hundreds of millions of records published in common formats and aggregated into centralised portals. Nevertheless, new challenges emerged from this initiative for effectively using such a large volume of data. In particular, as the number of species, geographic regions, and institutions continue growing, answering questions about the complex interrelationships among these data become increasingly difficult. The Semantic Web (SW) [6] provides possible solutions to these problems by enabling the Web of Linked Data (LD) [7], where data objects are uniquely identified and the relationships among them are explicitly defined. LD is a powerful and compelling

approach for spreading and consuming scientific data. It involves publishing, sharing and connecting data on the Web, and offers a new way of data integration and interoperability. The driving force to implement LD spaces is the RDF technology. Moreover, there is an increasing recognition of the advantages of LD technologies in the life sciences [8,9].

In this same direction, CENPAT[1] has started to publicly share its data under Open Data licence.[2] Data are available as Darwin Core Archive (DwC-A) [10], which are a set of files for describing the structure and relationships of the raw data along with metadata files conforming the DwC standard. Nevertheless, the well-known IPT platform focuses on publishing content in unstructured or semi-structured formats but reducing the possibilities to interoperate with other datasets and make them accessible for machines. To enhance this approach, we present a transformation process to publish these data as RDF datasets. This process uses OpenRefine [11] for generating RDF triples from semi-structured data and define URIs. It also uses GraphDB [12], previously known as OWLIM [12], for storing, browsing, accessing and linking data with external RDF datasets. Along this process, we follow the stages defined in the LOD Life-Cycle proposed in [13]. We claim that this work is an opportunity to exploit data from biodiversity in Argentina because they had been never published as LOD.

This work is structured as follows. Section 2 describes the main features of the datasets selected and their relationships with DwC. Section 3 describes the transformation process to RDF, while section 4 presents its publication and its access. Section 5 shows the framework to discover links to other datasets. Next, section 6 presents the exploitation of the dataset. Finally, we draw conclusions and suggest some future improvements.

2   CENPAT Data Sources

In this section, before describing our datasets, we briefly explain the DwC standard and DwC-A, which these datasets are based on.

2.1   Darwin Core Terms and Darwin Core Archive

DwC [4] is a body of standards for biodiversity informatics. It provides stable terms and vocabularies for sharing biodiversity data. DwC is maintained by TDWG[3] (Biodiversity Information Standards, formerly The International Working Group on Taxonomic Databases). Its terms are organised into nine categories (often referred to as *classes*), six of which cover broad aspects of the biodiversity domain. *Occurrence* refers to existence of an organism at both particular place and time. *Location* is the place where the organism were observed (normally a geographical region or place). *Event* is the relationship between *Occurrence* and *Location* and register protocols and methods, dates, time and field notes.

---

[1] http://www.cenpat-conicet.gob.ar/
[2] https://creativecommons.org/licenses/by/4.0/legalcode
[3] http://www.tdwg.org/

Finally, *Taxon* refers to scientific names, vernacular names, etc. of the organism observed. The remaining categories cover relationships to other resources, measurements, and generic information about records. DwC also makes use of Dublin Core terms [14], for example: *type, modified, language, rights, rightsHolder, accessRights, bibliographicCitation, references.*

In the same direction, Darwin Core Archive (DwC-A) [10] is a biodiversity informatics data standard that makes use of the DwC terms to produce a single, self-contained dataset and thus sharing both species-level (taxonomic) and species-occurrence data. Moreover, each DwC-A includes these files. Firstly, the `core data file` (mandatory) consists of a standard set of DwC terms together with the raw data. This file is formatted as fielded text, where data records are expressed as rows of text, and data elements (columns) are separated with a standard delimiter such as a tab or comma. Its first row specifies the headers for each column. Secondly, the `descriptor metafile` defines how the core data file is organised and maps each data column to a corresponding DwC term. Lastly, the `resource metadata` provides information about the dataset itself such as its description (abstract), agents responsible for authorship, publication and documentation, bibliographic and citation information, collection method, among others.

2.2 Dataset Features

The datasets analysed belong to CENPAT and are available as DwC-A in an IPT server from this institution. They include collections of marine, terrestrial, parasites and plant species mainly registered from several points of the Argentinian Patagonia. Data are generated in different ways: some of them by means of electronic devices placed in different animals to study environmental variables, while others are observations of species in their natural habitat or species studied in laboratories. To ensure the quality of these data, the records have been structured according to the procedure described in [15].

Up to May 2017, CENPAT owns 33 datasets representing about 273.419 occurrence records, where 80% of them have been also georeferenced. Some of these collections contain unique data never published because of the age of the records (1970s). As a consequence, making this information available as LOD is so important for researchers, who are studying species conservation and the impact of man in biodiversity along the last years [16,17].

3 Linked Data Creation

Publishing data as LD involves data cleaning, mapping and conversion processes from DwC-A to RDF triples. The architecture of such a process is shown in Fig. 1 and has been structured as described in the following subsections.
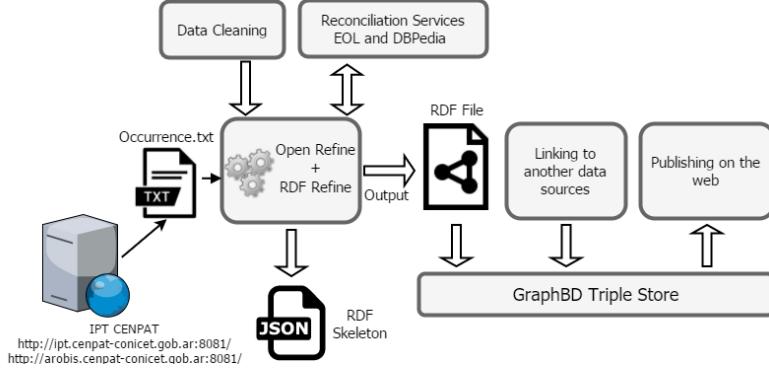
Figure 1. Transformation process for converting biodiversity datasets

3.1 Data Extraction, Cleaning and Reconciliation Process

The DwC-A are manually extracted from the IPT repository and their occurrences files (`occurrence.txt`) are processed using OpenRefine tool [11]. There, occurrences are cleaned and converted to standardised data types such as dates, numerical values, etc. and empty columns are removed. OpenRefine also allows adding reconciliation services based on SPARQL endpoints, which return candidate resources from external datasets to be matched to fields in the local datasets. In our process, we use DBpedia [18] endpoint[4] to reconcile the `Country` column with the `dbo:country` resource in DBpedia, the link between the resources is made through the property `owl:sameAs`. After that, if the reconciliation is done, we create a new column for the corresponding URI of the resource. In particular, we add the column named `dbpediaCountryURI` for the original `Country`.

Another reconciliation service[5] used, it was based on a taxonomic database *Encyclopedia of Life* (EOL)[6] which allows to reconcile *accepted names* in EOL database. Specifically, the reconciliation is applied to the column `scientificName` so that we create a new column named `EOL_page` for the EOL page describing the specie. Unfortunately, this whole process is time-consuming because not all values are automatically matched and thus ambiguous suggestions must be fixed. Moreover, in this phase only two columns have been possible to reconcile because the process returns unsuitable results using DBpedia services some columns like `institutionCode` or `locality`.

---

[4] https://dbpedia.org/sparql

[5] http://iphylo.org/~rpage/phyloinformatics/services/reconciliation_eol.php

[6] http://www.eol.org/

3.2 RDF Schema Alignment and URI Definition

After cleaning and reconciling, data are converted to RDF triples using RDF Refine[7], which is an extension of OpenRefine tool. RDF Refine allows users to go through a graphical interface describing the RDF scheme alignment skeleton to be shared among different datasets. The RDF skeleton specifies the subject, predicate and the object of the triples to be generated. The next step in the process is to set up prefixes. Since datasets include localities, locations and research institutes, we set up prefixes for well-known vocabularies such as the W3C Basic Geo ontology [19], Geonames [20], DBpedia, FOAF [21], Darwin-SW [22] for establishing relationships among DwC classes and Taxon Concept.[8] Table 1 shows the prefixes used.

Table 1. Prefix used in the mapping process.

Prefix	Description	URI
cnp-gilia	Base URI	http://crowd.fi.uncoma.edu.ar:3333/
dwc	Darwin Core	http://rs.tdwg.org/dwc/terms/
dws	Darwin-SW	http://purl.org/dsw/
foaf	Friend of a Friend	http://xmlns.com/foaf/0.1/
dc	Dublic Core	http://purl.org/dc/terms/
geo-pos	WGS84 lat/long vocab	http://www.w3.org/2003/01/geo/wgs84_pos#
geo-ont	GeoNames	http://www.geonames.org/ontology#
wd	Entitys in Wikidata	http://www.wikidata.org/entity/
wdt	Properties in Wikidata	http://www.wikidata.org/prop/direct/
txn	Taxon Concept Ontology	http://lod.taxonconcept.org/ontology/txn.owl#

In order to generate URI for each resource, in this approach we used GREL (General Refine Expression Language) also provided by OpenRefine, the general structure of the URIs is described below:

$$\texttt{http://[base\_uri]/[DwC class]/[value]}$$

where: `[base_uri]` is the one specifies in Table 1, `[DwC class]` is the respective DwC class and `[value]` is the value of the cells in the file of occurrences. It is also important to note that the generated URIs are instances of the classes defined in the DwC standard. Finally, the resulting RDF triple for an occurrence is:

```
SUBJECT: <base_uri/occurrence/f6bbf85d-85ea-4605-87fa-d81aca73a1cd>
PREDICATE: rdf:type
OBJECT: dwc:Occurrence
```

Table 2 describes the mapping performed and which columns have been used to generate the main URIs.

---

[7] http://refine.deri.ie/

[8] http://lod.taxonconcept.org/ontology/txn.owl

Table 2. The first part of the table shows the main classes corresponding to the categories of the DwC standard. Moreover, the columns of the DwC-A file used to generate URIs. The second part shows the properties used and an example of the literals obtained from the columns of the file of `occurrences.txt`. For simplicity, the table shows only the main properties, see the complete scheme at `https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/Open_refine_scripts/rdf_skelton.json`

Class	Columns used to create URI	URI example
dwc:Taxon	genus + specificEpithet	<base_uri:taxon/Mirounga_leonina>
dwc:Occurrence	id	<base_uri:occurrence/f6bbf85d-85ea-4605-87fa-d81aca73a1cd>
dwc:Event	id	<base_uri:event/f6bbf85d-85ea-4605-87fa-d81aca73a1cd>
dwc:Dataset	dataset	<base_uri:dataset/dwca-mamcenpat-v1.1>
dc:Location	id	<base_uri:location/f6bbf85d-85ea-4605-87fa-d81aca73a1cd>
foaf:Agent	institutionCode	<base_uri:agent/cenpat-conicet>

Property	Columns used	Example
dwc:class	class	"Mammalia"^^xsd:string
dwc:family	family	"Phocidae"^^xsd:string
dwc:genus	genus	"Mirounga"^^xsd:string
dwc:kingdom	kingdom	"Animalia"^^xsd:string
dwc:order	order	"Carnivora"^^xsd:string
dwc:phylum	phylum	"Chordata"^^xsd:string
dwc:scientificName	scientificName	"Mirounga leonina Linnaeus, 1758"^^xsd:string
txn:hasEOLPage	EOL_page	"http://eol.org/pages/328639"^^xsd:string
dwc:basisOfRecord	basisOfRecord	"PreservedSpecimen"^^xsd:string
dwc:occurrenceRemarks	occurrenceRemarks	"craneo completo"^^xsd:string
dwc:individualCount	individualCount	1^^xsd:int
dwc:CatalogNumber	CatalogNumber	"100751-1"^^xsd:string
geo-pos:lat	decimalLatitude	-42.53^^xsd:decimal
geo-pos:long	decimalLongitude	-63.6^^xsd:decimal
geo-ont:countryCode	country	"Argentina"^^xsd:string
dwc:verbatimEventDate	dwc:verbatimEventDate	"2004-10-22"^^xsd:date
foaf:name	recordedBy or InstitutionCode	"CENPAT-CONICET"@en.

4 Publishing and Accessing Data

The transformed biodiversity data have been published, and can to be accessed, through GraphDB. GraphDB is a highly efficient and robust graph database with RDF and SPARQL support. It allows users to explore the hierarchy of RDF classes (`Class hierarchy`), where each class can be browsed to explore its instances. Similarly, relationships among these classes also can be explored giving an overview about how many links exist between instances of the two classes (`Class relationship`). Each link is a RDF statement where its subject and object are class instances and its predicate is the link itself. Lastly, users also can explore resources providing URIs representing any of the subject, predicate or object of a triple (`View resource`).

Finally, Fig. 2 shows the resulting graph for the description of a *southern elephant seal skull*, which is part of the CENPAT collection of marine mammals and contains information about where has been found, who has been collected for, sex and scientific name, among others. Another way to access the same information is to explore the `View resource` in the GraphDB repository `http://crowd.fi.uncoma.edu.ar:3333/resource/find` for the specific occurrence `f6bbf85d-85ea-4605-87fa-d81aca73a1cd`, while the serialization of the complete graph in Turtle syntax can be consulted in.[9]
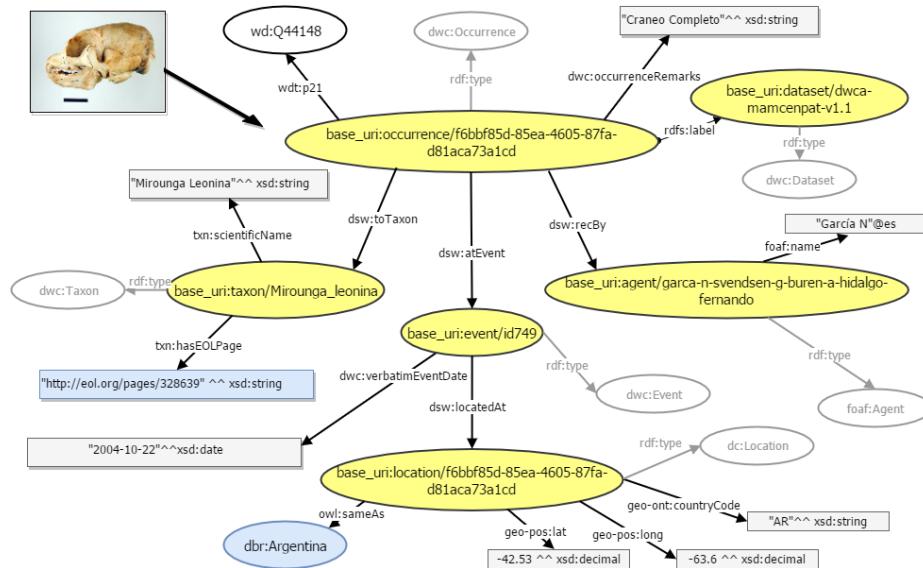

Figure 2. Figure shows links between instances of classes, `rdf:type` assertions are shown in light gray. In blue colour you can see the reconciled values .

---

[9] `https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/rdf/graph.ttl`, accessed at September 2017

5   Interlinking

Interlinking other datasets in a semi-automated way is crucial aiming at facilitating data integration. In this context, OpenRefine reconciliation service is able to match some links to DBpedia, but since it is still limited, our process should use more powerful tools to discover links to other datasets. For this task, our approach preliminarily integrate SILK framework[10] that uses `Silk-Link Specification Language` (Silk-LSL) to express heuristics for deciding whether a semantic relationship exists between two entities. For interlinking species between DBpedia and our dataset, we used *Levenshtein distance* a comparison operator that evaluates two inputs and computes the similarity based on a user-defined distance measure and a user-defined threshold. This comparator receives as input two strings `dbp:binomial` (Binomial nomenclature in DBpedia) and the combination of `dwc:genus` + `dwc:specificEpithet` (the concatenation of these two defines the scientific name of the species). The Levenshtein distance comparator was set up with `<Thresholds = "0.0" and Weight = "1">`. After the execution, SILK discovered 15 links to DBpedia with an accuracy of 100% and 85 link with an accuracy between 65% and 75%. In this case, we permit only one outgoing `owl:sameAs` link from each resource. The complete Silk-LSL script can be downloaded from.[11]

However, although a set of links has been successfully generated, users' feedback is needed to filter some species wrongly matched by the tool. Finally, we must identify further candidates for interlinking and tests other properties or classes from our dataset in order to increase the automatic capabilities of the framework.

6   Exploitation

This section shows how the different types of observations of species can be retrieved, complemented with information of another datasets and filtered by submitting SPARQL queries to GraphDB endpoint. Moreover, it provides some experiments in R by using the `SPARQL`[12] package. Each SPARQL query in following examples assumes the prefix defined in Table 1.

Total Number of Species in the CENPAT Dataset. The following query retrieves the species of the dataset. To this end, it includes the scientific name of the species and also its amount of occurrences, to execute this query in GraphDB see.[13] The Fig. 3 shows only the first resulting records.

---

[10] `http://silkframework.org/`

[11] `https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/SILK/`
  `link-spec.xml`, accessed at September 2017

[12] `https://cran.rproject.org/web/packages/SPARQL/SPARQL.pdf`

[13] `http://crowd.fi.uncoma.edu.ar:3333/sparql?savedQueryName=species-count`

```
SELECT ?scname (COUNT(?s) AS ?observations)
      {?s a dwc:Occurrence.
       ?s dsw:toTaxon ?taxon.
       ?taxon dwc:scientificName ?scname }
GROUP BY ?scname
ORDER BY DESC(COUNT(?s))
```

```
-------------------------------------------------------
scname                                      observations
-------------------------------------------------------
Otaria flavescens                               2,787
Mirounga leonina                                1,851
Merluccius hubbsi                                 368
```

Figure 3. Occurrences of each species that contains the dataset.

Occurrences by Year. The following query allows to observe the temporality of the occurrences and its results are visualised using R as shown the Fig. 4. The R script is available in.[14]

```
SELECT ?year (COUNT(?s) as ?count)
      {?s a dwc:Event.
       ?s dwc:verbatimEventDate ?date }
GROUP BY (year(?date) AS ?year)
ORDER BY ASC(?year)
```
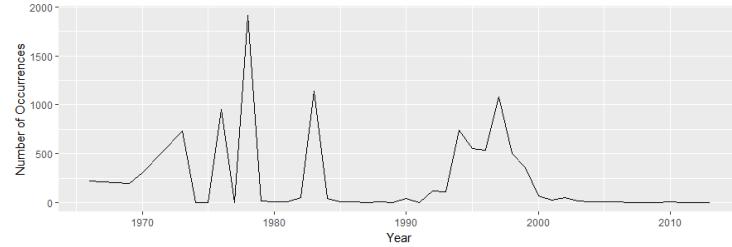

Figure 4. Simple plot using SPARQL and ggplot2 package for R.

Conservation Status of Species. Conservation status are defined by The IUCN Global Species Programme[15] and are taken as a global reference. Information about the state of conservation is missing in CENPAT datasets so that

---

[14] https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/
r-scripts/occurrences-by-year.R, accessed at September 2017
[15] http://www.iucnredlist.org/

providing these data linking other RDF datasets is highly desirable. To this end, the following query capture these missing data using the `owl:sameAs` property. The results are shown in Fig. 5, to execute this query in GraphDB, see.[16]

```
SELECT ?scname ?eol_page ?c_status
WHERE { ?s a dwc:Taxon.
        ?s dwc:scientificName ?scname.
        ?s txn:hasEOLPage ?eol_page.
        ?s owl:sameAs ?resource .
 SERVICE <http://dbpedia.org/sparql> {
        ?resource dbo:conservationStatus ?c_status.}
}
```

```
==========================================================================
scname                                          eol_page               c_status
--------------------------------------------------------------------------
Otaria flavescens                               http://eol.org/pages/328614    LC
Mirounga leonina                                http://eol.org/pages/328639    LC
Hydrurga leptonyx Blainville, 1820              http://eol.org/pages/328637    LC
Arctocephalus australis                         http://eol.org/pages/328623    LC
Lagenorhynchus obscurus Gray, 1828              http://eol.org/pages/317317    DD
Phocoena dioptrica Lahille, 1912                http://eol.org/pages/328461    DD
Hyperoodon planifrons Flower, 1882              http://eol.org/pages/328557    LC
Mesoplodon grayi Von Haast, 1876                http://eol.org/pages/328562    DD
```

Figure 5. Conservation status associated to the species: LC (Least Concern), DD (Data Deficient), EN (Endangered), VU (Vulnerable).

Locations of Marine Mammals. The last query is to retrieve the locations (latitude and longitude) for the species *Mirounga Leonina*. The results are depicted in Fig. 6 using R, and the script is available in.[17]

```
SELECT ?lat ?long
WHERE { ?s a dwc:Occurrence.
        ?s dsw:toTaxon ?taxon.
        ?taxon dwc:scientificName ?s_name.
        ?s dsw:atEvent ?event.
        ?event dsw:locatedAt ?loc.
        ?loc geo-pos:lat  ?lat .
        ?loc geo-pos:long ?long
FILTER (?lat >= "-58.4046"^^xsd:decimal && ?lat <= "-32.4483"^^xsd:decimal)
FILTER (?long >= "-69.6095"^^xsd:decimal && ?long <= "-52.631"^^xsd:decimal)
FILTER regex ( STR (?s_name ), "Mirounga␣leonina")}
```

7 Conclusions and Further Works

In this work we have presented a framework to publish biodiversity data from Argentinian Patagonia as LOD, which have initially been made available as

---

[16] http://crowd.fi.uncoma.edu.ar:3333/sparql?savedQueryName=conservation-status

[17] https://github.com/cenpat-gilia/CENPAT-GILIA-LOD/blob/master/r-scripts/positions-ml.R, accessed at September 2017
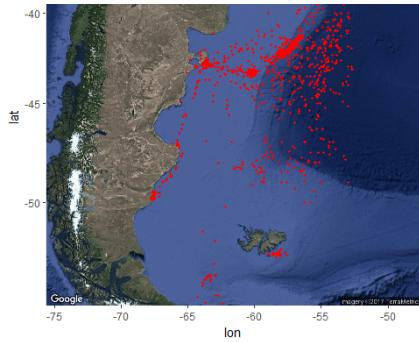
Figure 6. Visualization of animal movements using R

Darwin Core Archive files. The aim is to facilitate the access of researchers to important data and thus giving a valuable support to the scientific analysis of the biodiversity. In addition, this work is the first Argentinian initiative to convert biodiversity data according to the criteria established by LOD.

We have detailed the transformation process and explained how to access and exploit them, promoting integration with other repositories. Moreover, we have depicted this process using queries extracted from the domain of application. Such RDF repository is hosted at `http://crowd.fi.uncoma.edu.ar:3333/` together with an SPARQL endpoint, in this initial stage we store 202.119 triples.

As future works, we plan to automate some tasks of the process and interlink with more datasets. Moreover, providing easier SPARQL access for non-skilled users. Finally, we are analysing other ontologies such as ENVO [23], NCBI [24] and OWL Time [25] and working on a suite of complementary ontologies for describing every aspect of semantic biodiversity.

References

1. Craig Moritz, James L Patton, Chris J Conroy, Juan L Parra, Gary C White, and Steven R Beissinger. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. *Science*, 2008.
2. Adriana Vergés, Peter D Steinberg, Mark E Hay, Alistair GB Poore, Alexandra H Campbell, Enric Ballesteros, Kenneth L Heck, David J Booth, Melinda A Coleman, and Feary. The tropicalization of temperate marine ecosystems: climate-mediated changes in herbivory and community phase shifts. In *Proc. R. Soc. B*. The Royal Society, 2014.
3. Malcolm Scoble. Rationale and value of natural history collections digitisation. *Biodiversity Informatics*, 2010.
4. 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*, 2012.
5. Tim Robertson, Markus Döring, Robert Guralnick, David Bloom, John Wieczorek, Kyle Braak, Javier Otegui, Laura Russell, and Peter Desmet. The GBIF integrated

    publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PLoS One*, 2014.

6. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The Semantic Web. *Scientific American*, 2001.

7. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, 2009.

8. François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 2008.

9. Jouni Tuominen, Nina Laurenne, and Eero Hyvönen. *Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception.* Springer, 2011.

10. K Döring M Robertson T Remsen D, Braak. Darwin Core Archive How-To Guide. 2011.

11. Ruben Verborgh and Max De Wilde. *Using OpenRefine.* Packt Publishing Ltd, 2013.

12. Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, and Ruslan Velkov. OWLIM: A family of scalable semantic repositories. *Semantic Web*, 2011.

13. Sören Auer, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann, Michael Martin, Pablo N. Mendes, Bert Van Nuffelen, Claus Stadler, Sebastian Tramp, and Hugh Williams. Managing the Life-Cycle of Linked Data with the LOD2 Stack. In *International Semantic Web Conference (2)*, Lecture Notes in Computer Science, 2012.

14. Dublin Core Metadata Initiative et al. Dublin core metadata element set, version 1.1. 2012.

15. Mark J Costello and John Wieczorek. Best practice for biodiversity data management and publication. *Biological Conservation*, 2014.

16. Reed S Beaman and Nico Cellinese. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*, 2012.

17. Ana Vollmar, James Alexander Macklin, and Linda Ford. Natural history specimen digitization: challenges and concerns. *Biodiversity Informatics*, 2010.

18. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web*, 2007.

19. D Brickley. W3C Semantic Web Interest Group: Basic Geo (WGS84 lat/long) Vocabulary, 2011.

20. Marc Wick, B Vatant, and B Christophe. Geonames ontology. `http://www.geonames.org/ontology`, *accessed at Sep 2017*, 2015.

21. Dan Brickley and Libby Miller. The Friend Of A Friend (FOAF) vocabulary specification, 2007.

22. Steven J Baskauf and Campbell O Webb. Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. *Semantic Web*, 2016.

23. Pier Luigi Buttigieg, Evangelos Pafilis, Suzanna E. Lewis, Mark P. Schildhauer, Ramona L. Walls, and Christopher J. Mungall. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, 2016.

24. Scott Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 2012.

25. Time Ontology in OWL, 2006. `http://www.w3.org/TR/owl-time`, accessed at September 2017.

Enhancing the Discoverability and Interoperability of Multi-disciplinary Semantic Repositories

Doron Goldfarb[1][0000-0003-1183-6041] and Yann Le Franc[2][0000-0003-4631-418X]

[1] Environment Agency Austria, Vienna, Austria
[2] e-Science Data Factory, Paris, France
doron.goldfarb@umweltbundesamt.at, ylefranc@esciencefactory.com

Abstract. The aggregation of multi-disciplinary information is a challenge faced by large-scale data infrastructures serving scientific domains such as biodiversity, agronomy or ecology. This requires the integration of ontologies or thesauri from different domains. These semantic resources are often hosted within domain specific repositories which can be harvested for that purpose. The lack of discoverability, the technical and metadata heterogeneity of the semantic repositories pose a challenge for their effective integration. In this context, we argue that there is a need for a semantic lookup-service to access and use this heterogeneous landscape. We then present a proof-of-concept design and implementation for harvesting different ontology repositories (BioPortal, AgroPortal and EBI-OLS). We show some preliminary analytics and discuss technical issues regarding aggregation. Finally, we conclude with an open call for collaboration to address the issues hampering such initiatives.

Keywords: Ontology libraries, Semantic annotation, Ontology lookup service, EUDAT.

1 Introduction

Semantic technologies are increasingly used by domain-specific Research Infrastructures (RIs) and large-scale multi-disciplinary infrastructures such as EUDAT[1]. Semantically-enabled services offer a framework to aggregate data from multiple sources, enhancing discoverability and interoperability. The EUDAT pilot service B2Note[2] is one such service, allowing the creation of semantic annotations of datasets within and outside of the EUDAT infrastructure. The process of annotation is about "attach[ing] data to some other piece of data" [1]. In the scope of the Semantic Web, this usually refers to the contextualisation of information within a wider knowledge graph in order to support discovery and, eventually, automated reasoning. Such a vision can only be made possible through the wide-spread and repeated annotation with concepts defined in ontologies, thesauri or taxonomies. Throughout this work,

---

[1] http://www.eudat.eu/
[2] https://eudat.eu/news/annotate-your-research-data-with-b2note

we refer to such formalised knowledge representation structures as semantic resources, without any consideration for their format.

Providing domain specific concepts within a multi-disciplinary infrastructure requires their discovery and aggregation from different semantic resources available throughout the Web. This is also particularly true for RIs in the domain of biodiversity and ecology, where biology is linked to heterogeneous fields such as chemistry, molecular biology and earth science. In recent years, however, the number of available semantic resources has steadily grown to an extent making it hard to maintain the overview on "what's out there" and to identify the locations where they can be retrieved.

Dedicated repositories have thus been conceived to extend the discoverability of semantic resources by providing single access points for retrieving information about and from multiple, usually domain specific, semantic resources. Called "ontology libraries" by d'Aquin and Noy [2], these semantic repositories often provide programmatic access as (REST) API or via query languages such as SPARQL. They can thus be used to identify available semantic resources and moreover usually offer the advantage of hosting them in a homogenised form. This includes structured descriptive metadata about a resource such as name, acronym and version, as well as homogeneous extracts of its content which usually encompasses information about concepts and related properties.

Harvesting content from different domain specific semantic repositories can therefore support the aggregation of domain specific concepts for the semantic annotation of multi-disciplinary content. This endeavour, however, still remains a challenge as the large number of available semantic repositories raises the problem of their discoverability and interoperability.

In the context of EUDAT we designed a proof-of-concept service to aggregate multi-disciplinary semantic resources. This Semantic Lookup Service shall periodically retrieve the content from a set of registered semantic repositories and feed the results into a search index supporting concept discovery and auto-completion, used by the data annotation service B2Note. The development of such a centralised platform will increase the discoverability of the existing resources for the domain knowledge experts and for the growing eco-system of semantic tools, supporting the re-use of the semantic resources. Furthermore, the aggregation of content from large numbers of semantic repositories enables various types of analysis and metrics.

We argue that such a service will be of benefit especially to the life-sciences domain, since a huge proportion of existing repositories, such as BioPortal [3] and EBI-OLS [4], is rooted there, reflecting the already established tradition of using semantic resources. Providing a consolidated view on the semantic resource landscape present there will support related researchers but also foster the re-use of their resources in related domains such as agronomy, biodiversity and ecosystems research. In the context of the latter for example, initiatives such as the ILTER (International Long Term Ecosystem Research) network increasingly employ semantic resources such as the Environmental Thesaurus [5] for facilitating the harmonisation of heterogeneous data from its members [6]. The establishment of a repository dedicated to such resources is planned [7] and will augment already existing initiatives such as AgroPortal [8] for

the Agronomy/Agrology domain. The aggregation of these repositories provides the opportunity to identify cross-domain overlaps in terminology, potentially leading to mutual re-use and better cross-domain interoperability.

The remainder of this paper is structured as follows. Section 2 provides an overview on related work, identifies the key challenges and argues for the need for harmonisation between the existing solutions. It is then followed by the description of the design and first implementation of the proof-of-concept Semantic Lookup Service and our initial approach to harvesting concepts hosted in different semantic repositories in Section 3. Section 4 features a discussion of the results while section 5 gives an outlook for future work.

2   Related Work and Challenges

Semantic repositories seek to offer unique software platforms for extending the discoverability of semantic resources. Several implementations and approaches to semantic repositories have been developed and a first classification was proposed in 2012 by d'Aquin et al. [2]. While some of the approaches mentioned by the authors appear to have stalled, many others have emerged such as for example the SKOS oriented FINTO[3] service which is based on the SKOSMOS framework [9], the generic and curated Linked Open Vocabulary platform [10], the ANDS Research Vocabulary service[4] based on SISSVOC [11], and Ontobee [12]. In the biomedical domain, projects such as the above mentioned BioPortal and EBI-OLS evolved into advanced repositories. BioPortal has been reused in a growing variety of domains including Agriculture (AgroPortal) and Earth Sciences[5]. All these platforms offer means to harvest their content via RESTful APIs or SPARQL endpoints and thus support the aggregation of their content. The current variety of technical solutions increases the choice of offered functionalities but comes with burdens for interoperability. Indeed, the comparison of the different repositories revealed a large metadata and API heterogeneity. This represents a challenge to aggregate these resources into a multi-disciplinary semantic index.

Another major challenge is the discoverability of semantic resources. Indeed, the increasing number of semantic repositories makes it difficult to find all of them. In addition, many resources are not necessarily registered in a repository and can be rarely found via Google searches, leading to the situation that they are only known and (re-)used within a specific community.

We identified three main needs to address these challenges: a common metadata description, a common framework for API interoperability and a central hub to access the wealth of semantic resources.

The need for a common metadata description for semantic resources and semantic repositories has been identified by several initiatives such as the OBO Foundry [13]

---

[3] http://finto.fi/en/
[4] https://vocabs.ands.org.au/
[5] http://semanticportal.esipfed.org/

for the bio-medical domain, the ontology metadata schema proposed by LOV, and the Ontology Metadata Vocabulary [14]. One of the key challenges is to find a consensus between these different initiatives and to define a unique minimal common metadata set in order to enhance the interoperability between the different existing resources.

The problem of API interoperability is a generic problem for interlinking web services infrastructure. In the past few years, different initiatives have emerged to address this issue, including, for web-based APIs, the W3C HYDRA working community [15], the OpenAPI Initiatives[6] and the smart API [16], or, for RDF based datasets and resources, the W3C Vocabulary of Interlinked Datasets[7] and Linked Data Fragments[8]. The design of a central hub for API-providing repositories represents a unique opportunity to test and benchmark the different approaches to API interoperability.

The vision of a centralised service for discovering, searching, exploring and reusing semantic resources and related documents has already been proposed by several initiatives. Semantic search engines such as Swoogle [17], FalconS [18] or Watson [19] aimed at crawling and mining the web for semantic resources and offered means to search the results. Although they became valuable resources for knowledge workers, these different initiatives appear to have been discontinued. Other approaches sought to provide distributed search facilities across semantic repositories, such as the "Network of Ontology Repositories" [20], OntoCAT [21] and OntoHub [22].

In contrast to the latter, the approach presented in this paper aims at proposing a centrally aggregated search index which is not limited to locally stored resources but includes concept level extracts from remotely harvested semantic repositories. This index can then be used as a semantic search engine based on - in contrast to Swoogle and related work - registered resources and repositories instead of Web crawling.

We believe that such an approach is beneficial for the quality of the content and its re-use. Moreover, it would provide means for large-scale analysis of the different resources such as the recent analysis performed in BioPortal [23] and provide meaningful information to ontologists, data scientists and knowledge engineers.

3   The Semantic Look Up Service

Based on the need identified in the previous sections, we designed an initial proof-of-concept service for cataloguing semantic resources. In this section we first describe the general design principles, then the current implementation.

The design of the service followed several identified requirements, centred on providing means to adequately describe and register semantic repositories in a way that the relevant information about the hosted resources and the concepts therein could be mapped to a common representation for indexing. Stored in a database, such descriptions should enable a harvesting service to retrieve the content at regular intervals, transforming and storing it as common index representation. This basic infrastructure should be flexible enough to allow the provision of additional services such

---

[6] https://www.openapis.org/
[7] https://www.w3.org/TR/void/
[8] http://linkeddatafragments.org/

as a public catalogue of described repositories/resources and an API for data analysis. A schematic of the proposed architecture and the data flow between the different elements of the service is shown in Figure 1.

The service should be composed of 5 main components: (1) a web interface to capture the description of external semantic repositories provided by their managers, (2) a database storing the repository descriptions used by (3) an information harvester collecting the descriptions to harvest the repository contents via their API and store them into the database. This information will be used to build and update (4) an index of all the concepts contained in the different registered repository for fast retrieval and use for semantic services. Finally, another web interface (5) should be designed to discover, visualise and interact with the catalogue, providing views of registered semantic repositories and the ontologies and vocabularies harvested, including information that can be gathered and extracted from them, such as inter-repository overlap both on resource and on concept level.
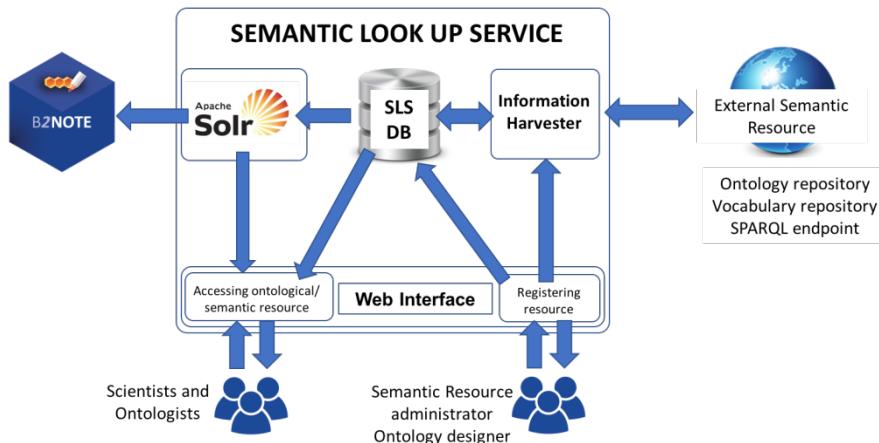

Fig. 1. Proposed architecture and data flow in EUDAT's Semantic Lookup Service

3.1 Current Implementation

We developed an initial implementation based on Python and command line scripts to provide an initial index of bio-medical semantic concepts for B2NOTE. We created custom scripts to harvest 5 million concepts from 494 ontologies hosted in BioPortal which populated a first instance of a SolR[9] index. The initial SolR schema has a focus on information about terms/classes defined/reused in the individual resources, but our approach can easily be extended to also cover other aspects such as properties. Table 1 lists the respective index fields. Besides fields describing concepts on individual and resource level, additional ones are conceived for filtering/organising search results. One such field is dedicated to listing concept reuse across resources and can support

---

[9] http://lucene.apache.org/solr/

the ranking of results, while another one provides information about the domain(s) the concept belongs to which can be used for limiting the search space. While the former can be automatically derived by analysing the harvested resources for conceptual overlaps, the latter should be provided by the repositories themselves and is currently only rarely available.

Table 1. Fields used for the SOLR lookup index

Concept IRI	IRI of the concept.
Concept Label	Human readable label of the concept.
Concept Description	Definition of the concept.
Concept Short_form	Short form of the concept.
Concept Synonyms	List of synonym labels referenced for the concept.
Resource Acronym	Acronym of the resource the concept pertains to.
Resource IRI	IRI of the resource the concept pertains to.
Resource name	Name of the resource the concept pertains to.
Resource vdate	Resource "released" field information.
Resource version	Resource "version" field information.
Acrs_of_resources_reusing_uri	List of acronyms for the resources reusing the concept.
Domains (not harvested yet)	Scientific domain covered by the resource

The initial workflow for harvesting the BioPortal API was directly coded as a Python script. This "plug-in" based approach is clearly not scalable and requires building scripts for every repository and maintaining them accordingly. We thus concentrated on developing a more efficient and generic approach for harvesting different repositories, focusing our effort on harvesting REST APIs and leaving the harvesting of SPARQL endpoints for future work.

To acquire the information needed for the SolR index (see Table 1), none of the analysed platforms, i.e. BioPortal, EBI-OLS and AgroPortal, provide one single function for retrieving the full set of fields and their harvesting thus involves several steps. We identified a two-step pattern to access this information. For each repository, an initial request retrieves basic resource level information and for each of the retrieved resources, additional requests then acquire information both on resource as well as on concept level.

Our initial approach to provide a common description framework for the three observed repositories uses a JSON description of the query sequence identified above. It contains information about the necessary query URLs as well as the locations of the data in the response sets from the different APIs, mapped to the respective fields of the SOLR index via JSONPath[10] expressions. Since a more detailed description is beyond the scope of this paper, it will be provided in a separate work. As shown in the next section, however, this approach enables us to successfully re-use one base implementation across three repositories, one of which having quite a different API implementation compared to the other two.

4   Analysing three Semantic Repositories

We applied our JSON/JSONPath based harvesting description to three existing repositories, BioPortal and its derivative AgroPortal, as well as EBI-OLS, this section provides results of a preliminary analysis. We were able to retrieve 96% of the available concepts (13,660,813 out of 14,226,183) from 93% of the semantic resources (786 out of 843). A first analysis of this initial dataset showed that 8,840,852 concepts were unique when distinguished by their URI and 6,109,756 when compared by strictly matching preferred labels. Table 2 provides data for each repository. Included in parentheses is the number of encountered resources having at least one concept vs. the stated number of resources hosted by each repository. The lower numbers for resources with concepts is due to the fact that missing ones are registered as private, as summary description only, feature only properties but no concepts, or are SKOS based resources. The latter - including a large fraction of AgroPortal, e.g. AGROVOC - were not harvested as they are considered as instances and not classes in the repository. We will investigate this in more detail and seek to extend our harvester in this regard.

Table 2. Total and unique number of concepts (no instances) present in three repositories

	AgroPortal (63/64)	BioPortal (534/586)	EBI-OLS (189/193)
Total	1,198,472/1,200,845	7,569,311/8,130,580	4,893,030/4,894,758
Unique URI	1,186,681	6,659,704	4,235,425
Unq. Label	1,122,242	5,379,485	3,938,468

Checking for inter-repository overlap on resource level, we compiled an alignment of the resource descriptions retrieved from each API. Noting some ambiguities regarding resource acronyms and discrepancies regarding resource URIs, we aligned the resources by their name. This operation still required some manual editing to compen-

---

sate for encountered naming discrepancies. The assessed resource overlap is presented in Figure 2, updating and extending the comparison between BioPortal and EBI-OLS as of 2011 [21]. A strong overlap between EBI-OLS (grown by about ⅓ since 2011) and BioPortal (Almost tripled since 2011) becomes immediately visible, their common 131 resources (113 of them OBO foundry related) now represent 67.9% of the EBI-OLS resources (45.5% in 2011). Another interesting observation is that on resource level, AgroPortal has higher overlap with EBI-OLS than with BioPortal. This is due to a set of crop specific ontologies taken from the Crop Ontology[11] project, hosted in both EBI-OLS and AgroPortal but not in BioPortal. Besides the identified overlaps, each repository provides a unique set of resources not present in the others.
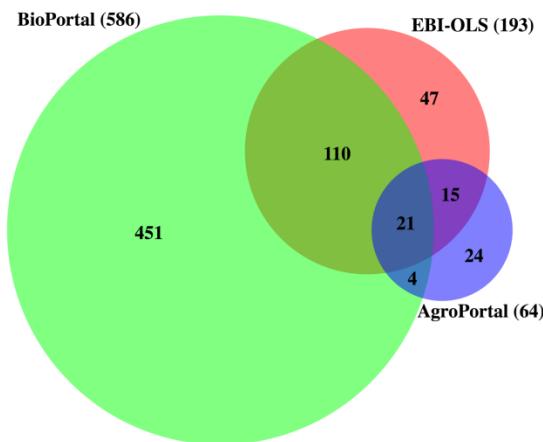

Fig. 2. Resources shared between EBI-OLS, BioPortal and AgroPortal

As stated above, the comparison of the resource descriptions revealed some relatively rare (affecting 36 out of 786 resources investigated) but nevertheless notable ambiguities regarding resource acronyms. We found (1) similar resources to have different acronyms in different repositories, such as the "Beta Cell Genomics Ontology" having "bcgo" in EBI-OLS and "obi_bcgo" in BioPortal, and (2) different resources to share similar acronyms across different repositories, such as "aeo" standing for "Anatomical Entity Ontology" in EBI-OLS/BioPortal and for "Agricultural Experiments Ontology" in AgroPortal.

Our observations show that acronyms are not always uniquely assigned to the same resources across repositories and such discrepancies can also be found in resource names. These ambiguities in our opinion provide a rather strong incentive to seek for a global name authority for semantic resources. We note, however, that establishing unique resource prefixes across different domains and communities could involve significant effort such as changing existing identifiers, which, as stated in [24], might outweigh the intended benefit. Given that we encountered relatively few such cases in

---

[11] http://www.cropontology.org

the observed repositories nevertheless suggests that the consultation of services such as prefixcommons[12] should be integrated in existing repositories.

For the further analysis of the aggregated resources, future work will concentrate on mappings at concept level, considering algorithms such as LOOM [25] and other approaches already employed in BioPortal[13] and other initiatives.

5 Conclusions & Outlook

In this paper we argue that there is a clear need for a centralised semantic look up service allowing to aggregate multi-disciplinary semantic resources. We emphasised that this effort is hampered by the lack of a common metadata set to describe the semantic resources, API interoperability and discoverability of the existing resources. We described our initial approach to build an index of multi-disciplinary concepts for semantically-enabled services in EUDAT. Our work presents the design of an initial proof-of-concept enabling different stakeholders to start referencing the different resources and serving as testbed for different solutions to aggregating multiple semantic repositories. We are presenting here our current implementation and the initial harvesting experiments that were performed. These experiments show that the centralised aggregation of multiple repositories also enables cross-repository analysis which is useful for studying the present landscape and improving data quality and thus interoperability.

In the future, we will extend the number of repositories to propose a general description of the harvesting workflows and we will align this work with the existing state-of-the art approaches for API interoperability. In parallel we will design and build an initial web interface to further extend the number of repositories and capture mapping information between their internal data model and the information needed for the SolR index. Finally, we will work on improving the SolR index by adding filters and facets to provide more usable search and exploration facilities across millions of terms.

We strongly believe that this effort can only be achieved through an extensive international collaboration between the different semantic repositories, the different initiatives proposing metadata representation of the semantic resources and the initiative working on API interoperability. Such collaboration has been discussed and initiated during different events organised in the context of EUDAT[14] and in collaboration with LifeWatch Italy[15]. Since these topics are in line with the general scope of the RDA Vocabulary and Semantic Service Interest Group[16] we are now working in the

---

[12] https://prefixcommons.org

[13] https://www.bioontology.org/wiki/index.php/BioPortal_Mappings

[14] https://www.eudat.eu/events/trainings/co-located-eudat-semantic-working-group-workshop-9th-rda-plenary-barcelona-3-4

[15] http://www.servicecentrelifewatch.eu/ontology-semantic-web-for-biodiversity-ecosystem-research

[16] https://www.rd-alliance.org/groups/vocabulary-services-interest-group.html

context of this interest group and hope to raise the interest on a global level and work in alignment with similar initiatives such as OntoHub and OntoCAT.

References

1. Oren, E., Möller, K., Scerri, S., et al.: What are semantic annotations?. DERI, Galway (2006).
2. d'Aquin, M. and Noy, N. F.: Where to publish and find ontologies? A survey of ontology libraries. Web Semantics: Science, Services and Agents on the World Wide Web, 11, pp. 96–111 (2012).
3. Whetzel, P. L., Noy, N. F. and Shah, N. H.: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research, 39(Web Server issue), (2011).
4. Jupp, S., Burdett, T., Leroy, C. and Parkinson, H.: A new Ontology Lookup Service at EMBL-EBI. In: Malone, J. et al. (eds.) Proceedings of SWAT4LS International Conference 2015, pp. 118-119, CEUR-WS, vol. 1546, ceur-ws.org, online (2015).
5. Schentz, H., Peterseil, J. and Bertrand, N.: EnvThes-interlinked thesaurus for long term ecological research, monitoring, and experiments. In: Page, B. et al (eds.) Proceedings of the 27th Int'l Conference on Environmental Informatics, Shaker, Herzogenrath (2013).
6. Vanderbilt, K. L., Lin, C.-C., Lu, S.-S., et al.: Fostering ecological data sharing: collaborations in the International Long Term Ecological Research Network, Ecosphere, 6(10), pp. 1–18 (2015).
7. Fiore, N., Magagna, B. and Goldfarb, D.: EcoSemanticPortal: facilitating discovery and interoperability of ontologies and thesauri in the ecological domain. In: Algergawy, A. et al. (eds.) Proc. of the 2nd International Workshop on Semantics for Biodiversity, (In Press).
8. Jonquet, C., Toulet, A., Arnaud, E., et al.: Reusing the NCBO BioPortal technology for agronomy to build AgroPortal. In: Jaiswal, P. et al. (eds.) Proceedings of the 7th International Conference on Biomedical Ontologies (ICBO-BioCreative), CEUR-WS, vol. 1747, ceur-ws.org, online (2016).
9. Suominen, O., Ylikotila, H., Pessala, S., et al.: Publishing SKOS vocabularies with Skosmos. Article manuscript submitted for review, http://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf, last accessed 2017/09/18.
10. Vandenbussche, P.-Y., Atemezing, G. A., Poveda-Villalón, M. and Vatant, B.: Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. Semantic Web, 8(3), pp. 437–452 (2017).
11. Cox, S., Yu, J. and Rankine, T.: SISSVoc: A Linked Data API for access to SKOS vocabularies. Semantic Web, 7(1), pp. 9-24 (2016).

12. Xiang, Z., Mungall, C., Ruttenberg, A. and He, Y.: Ontobee: A Linked Data Server and Browser for Ontology Terms. In: Bodenreider, O. et al. (eds.) Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO), CEUR-WS, vol. 833, pp. 279-281, CEUR-WS.org, online (2011).

13. Smith, B., Ashburner, M., Rosse, et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11), pp. 1251 - 1255 (2007).

14. Hartmann, J., Palma, R., Sure, Y., et al.: Ontology Metadata Vocabulary and Applications, In: Meersman R. et al. (eds.) On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops. OTM 2005. LNCS, vol. 3762, pp. 906-915, Springer, Heidelberg (2005).

15. Lanthaler, M. and Gütl, C.: Hydra: A Vocabulary for Hypermedia-Driven Web APIs. In: Bizer, C. et al. (eds.) Proceedings of the WWW2013 Workshop on Linked Data on the Web, CEUR-WS, vol. 996, ceur-ws.org, online (2013)

16. Verborgh, R. and Dumontier, M.: A Web API ecosystem through feature-based reuse. arXiv preprint, arXiv:1609.07108, arXiv (2016).

17. Finin, T., Ding, L., Pan, R., et al.: Swoogle: Searching for knowledge on the semantic web. In: Cohn, A. (ed.) Proceedings of the 20th National Conference on Artificial Intelligence, pp. 1682–1683, AAAI Press, Palo Alto (2005).

18. Cheng, G., Ge, W. and Qu, Y.: Falcons: searching and browsing entities on the semantic web. In: Huai, J., et al. (eds.) Proceedings of the 17th international conference on World Wide Web, pp. 1101–1102, ACM, New York (2008).

19. d'Aquin, M., Baldassare, C., Gridinoc, L., et al.: Watson: Supporting next generation semantic web applications. In: Nunes, M. P., et al. (eds.) Proceedings of the IADIS International Conference on WWW/Internet. IADIS Press, (2007).

20. Viljanen, K., Tuominen, J. and Hyvönen, E.: A Network of Ontology Repositories, https://seco.cs.aalto.fi/publications/2010/viljanen-et-al-onki-nor-2010.pdf (2010), last accessed 2017/07/28.

21. Adamusiak, T., Burdett T., Kurbatova, N., et al.: OntoCAT -- simple ontology search and integration in Java, R and REST/JavaScript, BMC Bioinformatics, 12, 218 (2011).

22. Mossakowski, T., Kutz, O. and Codescu, M.: Ontohub - a repository engine for heterogeneous ontologies and alignments, http://ontolog.cim3.net/file/work/OpenOntologyRepository/Ontohub/ontohub--TillMossakowski-et-al_20130621a.pdf, last accessed 2017/07/28.

23. Kamdar, M. R., Tudorache, T. and Musen, M. A.: A Systematic Analysis of Term Reuse and Term Overlap across Biomedical Ontologies. Semantic Web, (preprint), IOS Press (2017).

24. McMurry, J. A., Juty, N., Blomberg, N., et al.: Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biology, 15(6), (2017).

25. Ghazvinian A., Noy, N. F., and Musen, M. A.: Creating Mappings for Ontologies in Biomedi-cine: Simple Methods Work. In: AMIA Annual Symposium Proceedings, pp. 198-202, AMIA, online (2009).

Integrated Semantic Search on Structured and Unstructured Data in the ADOnIS System

Friederike Klan, Erik Faessler,Alsayed Algergawy, Birgitta König-Ries, and
Udo Hahn

Friedrich-Schiller-Universität Jena, Jena, Germany
`firstname.lastname@uni-jena.de`

Abstract. We introduce ADOnIS, an information system which coherently integrates two important, yet mostly disparate data sources, namely structured, tabular data, and unstructured data in terms of publications. The integration is achieved by providing the underlying background knowledge of the domains involved in terms of adequately tailored ontologies. Once the two basic data sources are semantically linked, entirely novel opportunities for cross-source information retrieval arise which we will highlight in this paper.

1 Introduction

Two mutually separated "data cultures" have emerged over the years and still persist in the field of information systems. On the one hand, the database community focuses on the *structured* representation of slices of the reality, typically in terms of relations and tables. On the other hand, the information retrieval community deals with, from a computational view, *unstructured* data, namely documents as streams of characters (and other media types, such as visual data) and tries to computationally interpret (and thus restructure) the meaning encoded in these textual data carriers. Both worlds rest on solid mathematical foundations and stable technical implementations on the basis of which huge amounts of structured and unstructured data can be managed and searched on an industrial scale. Yet, with the exception of activities aiming at the *Semantic Web* (for a survey, cf. [20]) they currently lack crossover.

This lack of integration hampers the usability of data at all levels. Consider, as a concrete example, an interdisciplinary research community such as the one established in the collaborative research center (CRC) AQUADIVA, our research environment [16].[1] AQUADIVA explores the role of water (Aqua) and biodiversity (Diva) for shaping the structure, properties and functions of the earth's subsurface. When a graduate student enters the CRC, she might be interested in the transport of viruses in the geological subsurface. In order to get started the student searches for an overview of the state of the art and hints what has been done on this topic in AQUADIVA so far. So she searches for relevant publications in portals like PUBMED or GOOGLE SCHOLAR and poses search queries

---

[1] `http://www.aquadiva.uni-jena.de/`

to the BExIS 2 data portal, the central information system hub of the project to obtain data that have been collected already. Typically, the student will start with one query and then try to navigate results and find related entries.

Her success will strongly depend on her familiarity with the special mix of domains, skills of interacting with search engines and data repositories (including SQL/SPARQL-style query languages), her knowledge of linguistic variants and the taxonomic structures of the relevant sublanguages. For instance, queries for "virus transport subsurface", "virus transport soil", and "phages transport soil" typically return only partially overlapping result sets in PubMed or standard data management systems. This is due to simplistic string matching criteria, the incapability to account for linguistic variations of the same content (inflection variants, phrasal paraphrases, or synonyms) and the general lack of conceptual background knowledge (e.g., the taxonomic or partonomic structure of the domains' terminologies).

In our work, we aim to account for these deficiencies in a systematic way. The solution we propose is implemented in ADOnIS, the *AquaDiva Ontology-based Information System* that provides integrated and seamless access to structured data and unstructured publications by making use of a variety of semantic technologies such as ontologies and natural language processing (NLP) tools. With this, we hope to reduce the cognitive burden put on searchers while, at the same time, we intend to increase the coverage and quality of search results. In this paper, we briefly describe the methodologies underlying ADOnIS and the way users can interact with the system.

2 Related Work

Data in general and scientific data specifically can be roughly categorized into structured and unstructured data. Unstructured data has no predefined data model and is typically text-heavy. Due to its unstructured nature, it is a challenging task to extract specific and useful information [6, 10]. Retrieval algorithms for unstructured data often rely on keyword-based indexing and comparison techniques. They typically offer a search box query interface, where the searcher can input keywords of interest. Due to its simplicity, this kind of user interface, is very intuitive and easy to use. This comes at a cost. The semantics of the search query in terms of a set of input terms is not explicitly given and needs to be revealed by the information system.

On the other hand, structured data is data that is organized according to a predefined (but not necessarily explicitly known) data model, such as a table in a relational database (known data model), a document in RDF format ((partially) known data model) or a spreadsheet (unknown implicit data model). This predefined data model (if known[2]) enables search based on structured queries (e.g. SQL or SPARQL queries) with a well-known semantics. Although these kind

---

[2] In cases where the underlying data model is implicit (e.g. in spreadsheets), it needs to be provided by the data creator or has to be automatically extracted using machine-learning techniques. The latter can be particularly challenging, since in contrast to

of query interfaces make it easy to effectively identify and discover a piece of information and access it in concise way, they are rather complex and thus less suited to users with a non computer science background. Recent approaches have therefore started to combine and integrate *keyword-based* search approaches for unstructured data and *concept-based* approaches for structured data [3, 6, 2, 18, 19].

K-search is one of the earliest works on hybrid search that supports the retrieval of documents and knowledge [2]. The *K-search* approach aims at searching the Semantic Web as a collection of documents (unstructured data) and metadata (structured data). To achieve this goal, a hybrid strategy is proposed, where *keyword-based* and *metadata-based* search strategies are combined. *K-Search* uses two separate indexes for the hybrid search and combines the results afterwards via result intersection [10]. An ontology-based retrieval system is proposed in [6]. It adapts the classical vector space representation to be suitable for large-scale information sources. An ontology-based scheme is used to semi-automatically produce document annotations that are used for a semantic search. To cope with incomplete information in the knowledge base, the semantic search is combined with a conventional keyword-based search. Gärtner et al. [10] suggest a semantic search system (HS^3) that aims at semantically bridging the gap between structured and unstructured data. HS^3 is an automated system that augments an arbitrary knowledge base with additional information extracted from the Web. These information can then be used to build a document corpus and a combined index. This index is leveraged for a hybrid semantic search strategy that combines keyword-based and concept-based search. *TextTile* is a data visualization tool for datasets and query examination that requires a flexible analysis of structured data and unstructured text [9]. The tool includes a set of operations that can be interchangeably applied to structured as well as to unstructured textual data parts to generate useful data summaries. The tool does not make use of ontologies and semantic reasoning during the search process.

An semantic search architecture specifically designed for biodiversity data is suggested in [1]. The proposed system aims at improving the quality of the search results by exploiting ontologies and the contextual meaning of data. A mapping component links biodiversity data and concepts of a domain-specific ontology, *OntoBio*. A web interface supports end users to access data via SPARQL endpoints. In order to achieve this, the tool transforms domain ontologies, taxonomic information as well as biodiversity data into a common format. This has two disadvantages: datasets are duplicated and it becomes harder to reason on such big data. The *ELSEWEB* framework [22] aims at facilitating the integration of environmental data and providing semantic bridges between these data and species distribution models.

---

text-based documents, e.g. data tables, often reveal only scarce information that might give a hint to its meaning.

3 Overview of ADOnIS

We have implemented ADOnIS as an extension to the BExIS 2 data management platform [7][3]. In the following, we describe its two basic subsystems, namely the one dealing with already structured, tabular data (Sect. 3.1), and the one dealing with unstructured textual input on the basis of the semantic document search engine SeMedico (Sect. 3.2). The two components are supplemented by a graphical user interface that allows users to enter search terms based on which ADOnIS retrieves relevant data stored in BExIS 2 as well as publications (Sect. 4). A comprehensive view of the whole architecture of ADOnIS is provided in Fig. 1 which will be explained in the subsections to follow.

3.1 Handling Structured Data

Scientific data stored in BExIS 2 typically refer to field observations and measurements and are organized in tables. Each table and its corresponding meta information is referred to as a *dataset*. In addition to the data table containing the data values, each dataset comprises the *table schema* (name, datatype and unit of measurement for each data column) and *metadata* such as information about the data provider. Both, the actual data values and the table schema, are stored in a relational database.

To make the semantics of datasets explicit, we annotate each data table with conceptual knowledge encoded in ADOn , a domain-specific ontology expressed in Owl 2.[4] The ontology is tailored to the needs of the description of observational data from the life sciences domain. It only includes relevant classes and properties of these as TBox statements. Assertions about data values and data annotations, i.e. ABox statements, are not materialized in the ontology. Instead, we use the ontology-based data access system Ontop [5]. Based on a given ontology and a set of mappings that relate class and property symbols in the ontology to SQL views over the data in the database, Ontop provides a virtual RDF graph that can be queried using Sparql. This avoids duplication of instance data (that already reside in the relational database) and allows for sound and complete query answering in LOGSPACE under the *OWL 2 QL* entailment regime.[5] In order to retrieve datasets relevant to a certain search query, we generate a set of proper Sparql queries from the user-provided keywords, thus removing the burden from the searcher to formulate queries using a formal query language.

ADOn Ontology & Semantic Annotation. As core ontology, we use a modified version of the *Extensible Observation Ontology* (Oboe) [17] (version 1.2) that provides classes and properties for the description of field observations and measurements. Sets of related observations are organized in `oboe:Observation`

---

[3] `http://bexis2.uni-jena.de/`

[4] `https://www.w3.org/TR/owl-syntax`

[5] `https://www.w3.org/TR/owl-profiles/\#OWL_2_QL`
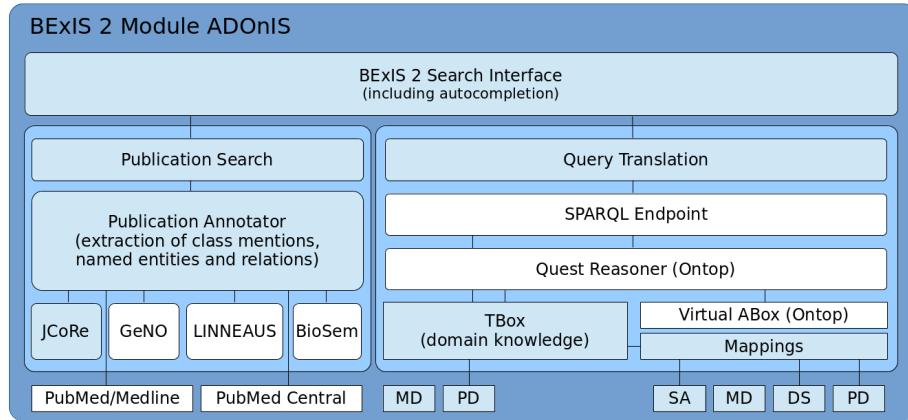
Fig. 1. System Architecture for ADOnIS

`Collection`s, which resemble the concept of a dataset in BExIS 2. Each data row in a BExIS 2 data table is modeled as one or more `oboe:Observation`s. An observation refers to an `oboe:Entity`, e.g. a *Tree*, and a set of `oboe:Measurement`s related to that entity. A measurement refers to an `oboe:Characteristic`, uses an `oboe:Standard` and results in a value. For instance, for a certain *Tree* entity, its *Circumference* (characteristic) might have been measured in meters (standard) and the measured value is 0.8. OBOE allows to indicate contextual relationships between observations, e.g. a tree might have been observed within a certain forest and this forest is located in a certain area. Modeling observations in this way enables logical inferences about entities and the relationships between them, as well as about measured characteristics of entities. In the life sciences domain, both observed entities and their characteristics are particularly important when trying to explain phenomena and thus play a key role when searching for datasets.

To cover domain-specific characteristics and entities, we reuse concepts from domain ontologies such as OBI (biomedical investigations),[6] ENVO (environmental features),[7] NCIT (biomedical concepts)[8] and CHEBI (chemical entities).[9] These were selected using the JOYCE tool for ontology selection and integrated into our ontology applying strict methodological criteria to guarantee non-redundancy, minimality, and optimal coverage [8]. These requirements were met by asserting subclass-relationships between concepts from a third-party ontology and either `oboe:Characteristic` or `oboe:Entity`. Since NCIT and CHEBI are huge in terms of the number of concepts they define, we used modularization techniques [8] to reuse only needed parts of these ontologies. We also

---

[6] `http://obi-ontology.org`

[7] `http://environmentontology.org`

[8] `https://evs.nci.nih.gov/`

[9] `https://www.ebi.ac.uk/chebi`

defined additional properties of `oboe:ObservationCollection`s, which directly relate datasets to observed entities, characteristics and standards (in contrast to OBOE, where these properties are related to individual observations). This enables efficient querying of these properties (instead of a potentially large set of observations (data rows) a much smaller number of datasets and their properties has to be inspected during search).

Each BExIS 2 data value/data column was (manually[10]) annotated with an ontology class corresponding to the entity it refers to, an ontology class modeling the characteristic that was measured and a class referring to the measurement standard that was used. Moreover, for each dataset, we indicated contextual relationships between the observed entities. The semantic annotations are stored in a relational database.

Ontop Mappings In order to enable SPARQL queries over the conceptual view given by the ontology, we defined mappings that relate BExIS 2 datasets, the entities and characteristics they refer to, the measured values and the dataset annotations residing in the relational database to class and property symbols in the ontology. These mappings are fixed for a given ontology and database. The subsequent mapping for example, creates a (virtual) instance for each characteristic measured in some annotated BExIS 2 dataset. It indicates the type of this instance (some subclass of `oboe:Characteristic`) as given by the semantic annotation stored in the database table `annotation` (cf. mapping below), and relates it to dataset instances that refer to this characteristic (not depicted).

```
mappingId CHARACTERISTIC-TYPE
target :crct_{crct_id} a <{crct}> .
source SELECT DISTINCT crct, chrct_id FROM annotation
```

Query Generation Using this approach, we can pose SPARQL queries about observational data stored in BExIS 2 on the schema level as well as on the level of individual data values. At the moment, we do not use the full potential of this solution, but rather restrict ourselves to the retrieval of BExIS 2 datasets based on keyword queries. For that purpose, we translate the search terms into a set of SPARQL queries. For each keyword that can be mapped to the label (via string comparison) of an ontology class C that is a subclass of `oboe:Characteristic`, we create the following SPARQL query (prefixes omitted) that returns all datasets that measure C.

```
SELECT DISTINCT ?dset
WHERE {
```

---

[10] We are currently working on a data upload wizard which analyzes new datasets to (semi-)automatically identify semantically annotated data attributes (the type of measurement referred to in a dataset column, its datatype and unit of measurement) that are already known to and maintained by ADONIS . Such a mechanism will enable semantic annotation with little user interaction.

```
    ?dset ad:refersToCharacteristic ?char.
    ?char a <URI of C> }
```

For each keyword that can be mapped to the label of an ontology class E that is a subclass of `oboe:Entity`, this is done in a similar way, which also accounts for contextual relationships between entities. We create a SPARQL query that asks for all datasets referring to entities of type E or to some entity that appears in the context of an entity of type E.

```
SELECT DISTINCT ?dset
WHERE {
    ?dset ad:refersToEntity ?ent.
    { ?ent a <URI of D> } UNION
    { ?ent ad:hasEntityContext ?entC.
      ?entC a <URI of D> } }
```

If the label of a characteristic was entered directly before the label of an entity in the search box, we interpret this as a search for the given characteristic measured for the given entity. In case a keyword neither matches the label of an `oboe:Characteristic` nor the label of an `oboe:Entity`, we search for datasets containing data values matching the keyword. Finally, we return the union of the resulting datasets. The required information about the type of each provided keyword is delivered by an autocomplete function that provides suggestions while the user is typing words in the BExIS 2 search box. The suggestions are generated based on an index of entity and characteristic class labels defined in the underlying ontology. The keywords provided by the user as well as the keyword-related information are passed to the structured search module, which has been implemented as web service with a REST-API.

3.2 Handling Unstructured Data

Unstructured data are handled by the SEMEDICO system which receives feeds from two sources, *viz.* more than 26 million life science abstracts from MEDLINE/PUBMED[11][12] and more than 1.5 million life science full texts from PUBMED CENTRAL from the open access subset. They are stored in a POSTGRESQL database.[13]

Ontologies & Semantic Annotation. Terminological and ontological resources for the indexing of all documents come from various sources. Most notable among them is the NCBI GENE database.[14] SEMEDICO's gene recognition and normalization engine maps gene mentions in the documents to unique NCBI

---

[11] `https://www.ncbi.nlm.nih.gov/pubmed`

[12] `https://www.nlm.nih.gov/databases/download/pubmed_medline.html`

[13] `https://www.postgresql.org/`

[14] `https://www.ncbi.nlm.nih.gov/gene`

Gene database entries to handle gene name synonymy and ambiguity. Additionally, SeMedico integrates the Gene Ontology (Go)[15] and the Gene Regulation Ontology (Gro)[16] for the semantic description of different types of gene events.

All resources are stored in a Neo4j[17] graph database for direct access to their hierarchical structure. All terminologies, ontologies and databases are converted into a common JSON format. This format is then imported into Neo4j using a custom Neo4j server plugin.

Natural Language Processing. Before Medline and PubMed Central documents are added to SeMedico's index, they undergo an extensive linguistic analysis. The goal is to identify textual units referring to gene/protein mentions, ontology concepts, gene interaction events and factuality markers for them as expressed in the documents. To be able to recognize such higher-level semantic concepts, it is necessary to do basic linguistic analysis first like sentence and token segmentation, part-of-speech tagging and chunking.

Semantic analysis includes species tagging by the Linnaeus tagger [11], gene mention tagging and normalization using GeNo [23], gene/protein event recognition with BioSem [4] and identification of event confidence ratings following the factuality rating as described by [13]. For BioSem, we use a model trained on the BioNLP Shared Task 2011 [15] training data that includes abstracts as well as full texts. MeSH, Go and Gro concepts are tagged by a dictionary component.

All documents undergo linguistic processing employing the Uima[18] component repository JCoRe [14]. The morpho-syntactic analysis includes the resolution of acronyms [21]. This step is crucial for the interactive disambiguation feature of SeMedico. We recognize textual mentions of ontology classes via preferred names and their synonyms. When searching, also subclasses of query concepts are automatically included in the search, leveraging the ontology's subclass hierarchy. Additionally, we employ dedicated named entity recognition tools for the detection of gene / protein mentions via GeNo [23] and species via the Linnaeus species tagger [11]. We also look for textually expressed relations between genes / proteins in publications. We employ BioSem [4] to extract mentions of gene / protein interactions from sentences such as

"Here we show that recombinant Pnc1 stimulates Sir2 HDAC activity."

were semantic connections between genes, proteins or, in this case, enzymes are described. Such relations have a high information value for researchers who look for interaction data on specific entities of interest. Modern relation extraction engines such as BioSem are far superior to simpler approaches which identify co-occurrences of entity within formal text units (e.g., sentences).

---

[15] http://www.geneontology.org/
[16] https://bioportal.bioontology.org/ontologies/GRO
[17] https://neo4j.com/
[18] https://uima.apache.org/

However, mere interaction extraction does not take into account the confidence level the authors of a publication assign to these observational data. Consider the following sentence: *"These results may suggest that mTOR-mediated autophagy inhibition may result in mesangial cell proliferation in IgAN."* While the sentence expresses some interaction between mTOR and igAN, the authors carefully use speculative words like *may* and *suggest*. Such information should be integrated into a scientific data portal to serve as an indicator how trustworthy an information item really is. We store all these annotations together with the original, raw documents in the document database.

In a last step, the analysis results required for semantic search are sent to an ELASTICSEARCH cluster for indexing. We use a custom ELASTICSEARCH plugin to have ELASTICSEARCH accept a term format that allows to exactly specify index terms within the ELASTICSEARCH index.

We model the publication search module as a web service disclosing a REST-like API. The API accepts parameters for a query string, a sorting criterion and the range of result documents that should be returned. The server then returns a JSON encoded response, including document text and bibliographic information.

4 Implementation & Preliminary Results

In this section, we introduce the GUI provided to the end user to facilitate the search process as well as preliminary evaluation results to demonstrate the effectiveness of the proposed method. To this end, we set up a running instance of the BExIS 2 system with the ADOnIS module that stores 55 real world datasets from the AQUADIVA project[19]. The datasets comprise 880 data columns and $539,774$ data rows in total. This results in $2,420,012$ single data values. For the unstructured data search results SEMEDICO stores more than 26M MEDLINE citations and approximately 1.5M PUBMED CENTRAL full texts from the open access subset in its index.

ADOnIS comes with a graphical user interface for the semantic search (Fig. 2). It is divided into three parts: the search box (top), where the user can enter keyword queries (one or more keywords), a section displaying publications (unstructured data) relevant to the query (left) and the list of retrieved BExIS 2 datasets (structured data) (right). An exemplary search using the keywords `groundwater`, `concentration of` and `nitrate` is shown in Fig. 2. The search delivers datasets that refer to the entity groundwater or entities that have been observed in the context of groundwater and datasets where the concentration of (characteristic) nitrate (entity) was measured. On the left-hand side, relevant publications are listed.

To demonstrate the effectiveness of the search functionality of ADOnIS we compared its results to those of the original keyword-based search provided by

---

[19] Currently, a subset of 15 datasets including 146 data attributes has been semantically annotated.
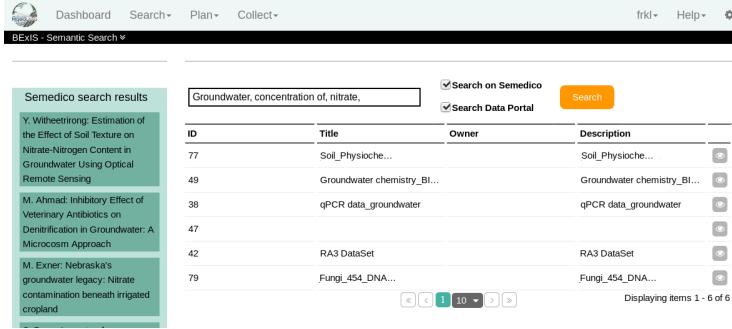
Fig. 2. Search Interface

BExIS 2 , which is powered by Apache Lucene[20] indexing both datasets and its accompanying metadata. As a preliminary evaluation, we've run the system with keyword queries relevant within the AQUADIVA project. We varied the query complexity by using one or more keywords. Exemplary results are reported in Table 1. In its current version, ADOnIS returns the union of both, the results returned by the semantic search and the results retrieved by the BExIS 2 standard search. This is to avoid an empty result set in cases where the semantic search does not retrieve any (exactly fitting) datasets. As a consequence, ADOnIS can just return additional datasets that have not been found by the original BExIS 2 search.

For a single keyword, ADOnIS and BExIS 2 typically return the same results, since those keywords are often explicitly mentioned either in the datasets itself or in the metadata. However, if we consider more complex queries, ADOnIS delivers relevant results that BExIS 2 does not discover. As a next step, we will extend this preliminary evaluation. In particular, we plan to invite formal feedback from the AQUADIVA researchers. This will cover both, an assessment of the relevance of the delivered search results[21] as well as an evaluation of the user interface. In addition, we will evaluate how well the search scales with an increasing number of datasets.

5   Conclusion

We introduced ADOnIS, an information system which coherently integrates two important, yet mostly disparate data sources, namely structured data from databases (or spreadsheets), on the one hand, and unstructured data in terms

---

[20] https://lucene.apache.org/

[21] Note that, even if datasets are annotated correctly, the search might deliver results that the user did not expect, since ADOnIS interprets the user's keywords in a certain way (cf. Sect. 3.1) that does not necessarily comply with the searcher's query intend. Such a mismatch would be discovered by a user study with the AQUADIVA researchers.

Table 1. Search results

Keywords	# of ADOnIS results	# of BExIS 2 results
RNA	16	16
soil moisture	6	2
chemical upper aquifer	2	0
groundwater concentration of nitrate	6	0

of publications, on the other hand. The integration is achieved by providing the underlying background knowledge of the domains involved in terms of adequately tailored ontologies. Once the two basic data sources are semantically linked, entirely novel opportunities for cross-source information retrieval arise.

6 Acknowledgments

References

1. F. K. Amanqui, K. J. A. Serique, S. D. Cardoso, J. L. C. dos Santos, A. C. F. Albuquerque, and D. A. Moreira. Improving biodiversity data retrieval through semantic search and ontologies. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, August 11-14, 2014 - Volume II*, pages 274–281, 2014.
2. R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi, and D. Petrelli. Hybrid search: Effectively combining keywords and semantic searches. In *5th European Semantic Web Conference, ESWC*, pages 554–568, 2008.
3. N. Bikakis, G. Giannopoulos, T. Dalamagas, and T. K. Sellis. Integrating keywords and semantics on document annotation and search. In *On the Move to Meaningful Internet Systems, OTM 2010 - Confederated International Conferences: CoopIS, IS, DOA and ODBASE, Hersonissos, Crete, Greece, October 25-29, 2010, Proceedings, Part II*, pages 921–938, 2010.
4. Q. C. Bui, E. M. van Mulligen, D. Campos, and J. A. Kors. A fast rule-based approach for biomedical event extraction. In *Proceedings of the BioNLP 2013 Shared Task Workshop*, pages 104–108, Sofia, Bulgaria, 2013.
5. D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, and G. Xiao. Ontop: Answering SPARQL queries over relational databases. *Semantic Web –Interoperability, Usability, Applicability*, 8(3):471–487, 2017.
6. P. Castells, M. Fernández, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. Knowl. Data Eng.*, 19(2):261–272, 2007.
7. J. Chamanara and B. König-Ries. A conceptual model for data management in the field of ecology. *Ecological Informatics*, 24:261–272, 2014.

8. E. Faessler, F. Klan, A. Algergawy, B. König-Ries, and U. Hahn. Selecting and tailoring ontologies with Joyce. In *Proc. of the Intl. Conf. on Knowledge Engineering and Knowledge Management*. Springer, 2017.

9. C. Felix, A. V. Pandey, and E. Bertini. Texttile: An interactive visualization tool for seamless exploratory analysis of structured data and unstructured text. *IEEE Trans. Vis. Comput. Graph.*, 23(1):161–170, 2017.

10. M. Gärtner, A. Rauber, and H. Berger. Bridging structured and unstructured data via hybrid semantic search and interactive ontology-enhanced query formulation. *Knowl. Inf. Syst.*, 41(3):761–792, 2014.

11. M. Gerner, G. Nenadic, and C. M. Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC Bioinformatics*, 11:85, 2010.

12. R. V. Guha, R. McCool, and E. Miller. Semantic search. In *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, May 20-24, 2003*, pages 700–709, 2003.

13. U. Hahn and C. Engelmann. Grounding epistemic modality in speakers' judgments. In D.-N. Pham and S.-B. Park, editors, *Trends in Artificial Intelligence. PRICAI 2014 –Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence. Gold Coast, Australia, 1-5 Dec, 2014*, number 8862 in Lecture Notes in Artificial Intelligence, pages 654–667. Springer, 2014.

14. U. Hahn, F. Matthies, E. Faessler, and J. Hellrich. UIMA-based JCoRe 2.0 goes GitHub and Maven Central: State-of-the-art software resource engineering and distribution of NLP pipelines. In *Proc. of the Intl. Conf. on Language Resources and Evaluation*, pages 2502–2509, Paris, 2016.

15. J. Kim, N. L. T. Nguyen, Y. Wang, J. Tsujii, T. Takagi, and A. Yonezawa. The genia event and protein coreference tasks of the bionlp shared task 2011. *BMC Bioinformatics*, 13(S-11):S1, 2012.

16. K. Küsel, K. U. Totsche, S. E. Trumbore, R. Lehmann, C. Steinhäuser, and M. Herrmann. How deep can surface signals be traced in the critical zone? merging biodiversity with biogeochemistry research in a central German Muschelkalk landscape. *frontiers in Earth Science*, 4:32, 2016.

17. J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279–296, Oct. 2007.

18. P. Peng, L. Zou, and Z. Qin. Answering top-k query combined keywords and structural queries on RDF graphs. *Inf. Syst.*, 67:19–35, 2017.

19. P. Peng, L. Zou, and D. Zhao. On the marriage of SPARQL and keywords. In *Web Technologies and Applications - 17th Asia-PacificWeb Conference, APWeb 2015, Guangzhou, China, September 18-20, 2015, Proceedings*, pages 3–16, 2015.

20. P. Ristoski and H. Paulheim. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 36:1–22, 2016.

21. A. S. Schwartz and M. A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003. Kauai, Hawaii, USA, January 3-7, 2003*, pages 451–462, 2003.

22. N. Villanueva-Rosales, N. R. D. Rio, D. Pennington, and L. G. Chavira. Semantic bridges for biodiversity sciences. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II*, pages 310–317, 2015.

23. J. Wermter, K. Tomanek, and U. Hahn. High-performance gene name normalization with geno. *Bioinformatics*, 25(6):815–821, 2009.

Terminologies as a neglected part of research data: Making supplementary research data available through the GFBio Terminology Service

David Fichtmüller[1], Maren Gleisberg[1], Naouel Karam[2], Claudia Müller-Birn[2], and Anton Güntsch[1]

[1] Botanic Garden and Botanical Museum (BGBM), Freie Universität Berlin, Germany
{d.fichtmueller@bgbm.de}
[2] Institute of Computer Science, Freie Universität Berlin, Germany
{naouel.karam@fu-berlin.de}

Abstract. In many research projects, much more data are created than made publicly available. Keeping research data deliberately closed or publishing only selected subsections of the gathered data are unfortunately common practices in academia. Fortunately, such problems have been getting more and more attention in the past years. However, another issue that is still often overlooked concerns research data that are generated as part of a research project but that are generally not considered part of the primary research data. One example for such neglected research data are terminologies such as controlled vocabularies that are used to describe or classify primary research data. In this paper we will outline the process that is used by the Terminology Service of the German Federation for Biological Data (GFBio) to prepare and process terminologies so that they can be included in the GFBio Terminology Service where they are made available to researchers within and outside the original research project. We will also show how making such supplementary research data publicly available will benefit the researchers who share them as well as the scientific community as a whole.

Keywords: GFBio, research data, terminology, ontology, terminology service

1 Introduction

In recent years, primary research data have been getting more attention as part of the publication process. Funding agencies such as the German Research Foundation (DFG[3]) and publishers are pushing scientists to publish the underlying research data along with the corresponding papers, or at least upload them to research data repositories. The DFG-funded project GFBio[4] (German Federation for Biological Data) is creating a dedicated repository for various kinds of biological research data and is developing supplementary tools for discovery and reuse of these data [2]. Various other initiatives are working on making research data publication and usage easier. One such initiative

---

[3] www.dfg.de
[4] www.gfbio.org
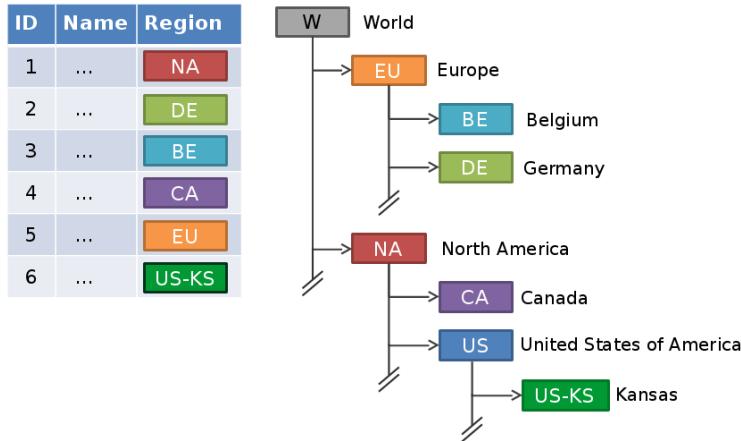
Fig. 1: A simple example of a geographic classification as it is used in research data and as it exists on its own with its definitions and connections.

is re3data.org[5] that has created an extensive registry of research data repositories, so scientists can easily find the repository that best fits their data and their needs. Another example is DataCite[6] who provides tools to make scientific data more citable and easier to find and reuse. Generally, the state of research data has significantly improved in recent years and will most likely continue to improve in the years to come. All of these tools and methods, however, generally only focus on the primary research data generated by the research projects. Another kind of data that is created during research projects is often overlooked: terminologies that are used to describe or classify records in the primary data. In scientific projects where several people are involved in the creation or gathering of the data, especially in large joint research cooperatives, it is vital to have a common understanding about the methods and categories used to describe these data. Ideally, this common understanding is expressed through written definitions of the terms prior to the collection of the first data. However, it is also possible that the conceptual agreement between the involved scientists was only achieved through ad-hoc discussions during data accumulation and was never formalized or documented. Even when common terms have been properly defined and documented, these documentations are often not published alongside the primary research data. This is a crucial loss of useful information, since definitions, synonyms and structural relations between terms usually cannot fully be extracted from the research data that is described using those terminologies, see Fig. 1.

Terminologies that are used to describe or classify primary research data can therefore

---

[5] www.re3data.org

[6] www.datacite.org

be considered as supplementary[7] research data, data that is not the primary focus of a research project, but vital to the accumulation of the primary research data.

2 Context and Related Work

2.1 What are Terminologies

In the context of the GFBio Terminology Service and this paper, a terminology is the overarching name for any set of fixed denotations that are used to describe something with the goal to reduce ambiguity and facilitate comparability. A terminology can range from a simple Controlled Vocabulary (a simple list of terms) to a complex Ontology (formal definitions of terms and their relations semantically expressed in a machine readable way). Terminologies can include translations and synonyms or aliases for individual terms.

2.2 What is GFBio and the GFBio Terminology Service

The German Federation for Biological Data (GFBio) is a national data infrastructure to store and facilitate access to biological and environmental research data. It offers services and resources to researchers for the archiving and publication of their research data as well as an open access portal to provide access to the data stored in the various data centers. The Terminology Service[8] (TS) of GFBio provides access to various terminologies for research data through one unified API [5]. Terminologies hosted at the TS can be distinguished into two groups: internal terminologies where data are locally stored and external terminologies[9] where the TS provides access to terminologies hosted on remove servers, examples for the latter case would be large databases like the Catalogue of Life (CoL)[10], the World Register of Marine Species (WoRMS)[11] or the GeoNames[12] Database. On the GFBio data portal, search queries for taxonomic names are extended using the TS to include synonyms and names of higher taxa, resulting in more relevant results for the users. The TS is therefore a vital component of the GFBio infrastructure. The GFBio Terminology Service can handle all kinds of terminologies, independent of their complexity, though the authors of terminologies to be included are required to at least provide definitions of the terms.

---

[7] Supplementary as in supplementary to the primary data, and not to be confused with supplementary data for journal publications where the supplementary refers to the primary research data being the supplement to the journal article.

[8] terminologies.gfbio.org

[9] In the context of this paper we will focus only on the preparation for terminologies to be imported as an internal terminology, as the process for connecting to an external terminology is completely different and beyond the scope of this paper.

[10] www.catalogueoflife.org

[11] www.marinespecies.org

[12] www.geonames.org

2.3 Related Initiatives

Different systems providing a comparable terminology service exist, the most widely used being Bioportal [7], a repository providing access to a large number of biomedical ontologies and Agroportal [4] its counterpart for agriculture and earth sciences. Finto (Finnish thesaurus and ontology service) [8] is a vocabulary service offering interfaces to ontologies from different domains, such as art, geography, science and medicine. The Ontology Lookup Service (OLS) [1] is a system integrating publicly available biomedical ontologies. And finally, Aber-OWL [3] is a framework that provides reasoning services over bio-ontologies. Specific project requirements motivated our decision of setting up our own solution, for instance regarding the range of heterogeneity of the considered terminologies or the necessity of combining ontology content with annotations to perform semantic search. More details about the requirements and a detailed comparison with existent systems can be found in [5].

3 Terminology Preparation Steps

If researchers want to have their terminology included in the GFBio Terminology Service, they need to contact the TS team, either directly or through the GFBio Submission Page[13]. To make a terminology fit the requirements for import in the Terminology Service, several processing steps might be required. These steps are done in close cooperation between the TS team and the scientist(s) providing the terminologies. The steps strongly vary between the individual terminologies, their type and complexity, and the additional work already provided by the involved scientists. The simplest case is when a dedicated list of terms is available as part of the supplementary research data, ideally with definitions and connections between the terms. In cases where no dedicated list of terms or formal documentation is available, the terms are extracted from the primary research data. This can range from simply exporting individual columns or tables from the set of the primary research data to doing complex parsing operations on the data to filter out the desired terminologies. The software used to do these extractions depends on the original data, e.g. when the terminology is included in the form of geographic data files, a common GIS software is used to extract it. The goal of the extraction process is to end up with a tabular file of the individual terms and their corresponding information, like hierarchies, if they can be extracted as well. Once the extraction is done, the scientists are asked to review the information for the completeness and correctness and provide any missing information that were not part of the original research data, such as definitions, translations or hierarchical structures in cases where they could not be extracted. The next step of the terminology processing is the data refinement and cleanup, which again is done in close contact with the contributing scientist(s). The refinement is usually done using OpenRefine[14], to catch errors like spelling mistakes in the term names, resulting in two very similar but not identical terms. Different additional tools are sometimes also

---

[13] https://www.gfbio.org/data/submit/generic; This is the same page as for the general GFBio data submission.

[14] www.openrefine.org

used to check for logical errors in the structure or other errors that cannot be checked using OpenRefine.

Each term of the terminology will get an individual URI which makes them addressable as a resource in the Semantic Web context. To avoid creating additional URIs for the same concepts, similar terminologies are searched for and if available, their terms are compared to the terms of the current terminology. In cases where terms are identical, the already existing URI is used. If terms are comparable but not identical to terms from other terminologies, then the relation between the terms is recorded by using properties such as skos:broader or skos:related. There are two options for contributing scientists if new URIs for the terms are assigned. The terms can either get the GFBio TS prefix[15], or they can provide their own prefix. The URIs with the TS prefix are resolvable and provide both human and machine readable formats depending on who is resolving the link. Custom URI prefixes on the other hand can help the branding of research projects, but the researchers are responsible for resolving the terms if they wish to have this highly recommended feature. In the end, the metadata of the terminology itself are formalized and the terminology is exported. Depending on its complexity this is usually SKOS, OWL or another RDF-based format which can then be imported into the GFBio Terminology Service. The export is done by creating a template in which the individual terms can be imported and using the OpenRefine templating engine to generate the final RDF file. After a final check and validation, the file is then imported into the GFBio TS, where the terms are then accessible via the TS API. When several scientists from the original research project wish to collaboratively and simultaneously work on reviewing and extending the terminology during the different feedback steps mentioned above, the TS team can provide dedicated tools.

4 Advantages of accessible Research Terminologies

There are several advantages that come with having research terminologies accessible. The foremost gain is that the primary research data itself becomes more understandable and reusable when the definitions and underlying hierarchies of the terms used to express it, are available as well. This is the primary use case of supplementary research data. These advantages can be further extended if the primary research data are served through a semantic aware search or portal, as this will allow for queries that also include synonyms or higher hierarchical terms, as demonstrated in [6]. Additional benefits could arise, if the primary research data not only uses the terms as a textual representation (i.e. copying its name) but as a semantic annotation, by using the concept URI to link to the term instead. Once the terms and their definitions are publicly available it strongly encourages their reuse. This could be in a subsequent project by the same researchers or even with researchers from other projects. Reusing terms not only saves time and effort for the people involved, but it makes the produced research data between the different projects more comparable, reusable and integrable. While journal publications of research papers and their subsequent number of citations are still the de facto standard to measure research impact, in recent years new approaches have come along to measure

---

[15] The TS URIs are formatted like this: http://terminologies.gfbio.org/terms/<terminology-name>/<term-name>

other kinds of scientific output as well, such as data publications or continuous work on service infrastructure. All terminologies on the GFBio Terminology Service can be cited as a research product which will give credit to the researchers who invested time and effort in creating them.

5   Conclusion

The GFBio Terminology Service is an important resource both for scientists who wish to share their terminologies that are used to describe and classify research data and for researchers who wish to apply existing terminologies and classifications to their own research data to improve their integrability. With reasonable additional effort the terminologies can be processed to be included in the TS and both the scientists who created the terminologies and the scientific community as a whole can benefit from this otherwise neglected research data.

References

1. R. G. Côté, P. Jones, R. Apweiler, and H. Hermjakob. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7(1):1–7, 2006.
2. M. Diepenbroek, F. O. Glöckner, P. Grobe, A. Güntsch, R. Huber, B. König-Ries, I. Kostadinov, J. Nieschulze, B. Seeger, R. Tolksdorf, and D. Triebel. Towards an integrated biodiversity and ecological research data management and archiving platform: The german federation for the curation of biological data (gfbio). In *44. Jahrestagung der Gesellschaft für Informatik, Stuttgart, Germany*.
3. R. Hoehndorf, L. Slater, P. N. Schofield, and G. V. Gkoutos. Aber-owl: a framework for ontology-based data access in biology. *BMC Bioinformatics*, 16(1):1–9, 2015.
4. C. Jonquet, A. Toulet, E. Arnaud, S. Aubin, E. Dzalé Yeumo, V. Emonet, V. Pesce, and P. Larmande. Reusing the NCBO BioPortal technology for agronomy to build AgroPortal. In *ICBO : International Conference on Biomedical Ontologies*, page 3 p., 2016.
5. N. Karam, C. Müller-Birn, M. Gleisberg, D. Fichtmüller, R. Tolksdorf, and A. Güntsch. A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data. *Datenbank-Spektrum*, 16(3):195–205, Nov 2016.
6. F. Löffler, K. Opasjumruskit, N. Karam, D. Fichtmüller, F. Klan, C. Müller-Birn, U. Schindler, and M. Diepnebroek. Honey bee versus apis mellifera: A semantic search for biological data. In E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, and O. Hartig, editors, *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I*, Lecture Notes in Computer Science, 2017.
7. N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M. D. Storey, C. G. Chute, and M. A. Musen. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web-Server-Issue):170–173, 2009.
8. O. Suominen, S. Pessala, J. Tuominen, M. Lappalainen, S. Nykyri, H. Ylikotila, M. Frosterus, and E. Hyvnen. Deploying national ontology services: From onki to finto. In *Proceedings of the Industry Track at the International Semantic Web Conference 2014*. CEUR Workshop Proceedings, October 2014.

What do Biodiversity Scholars Search for? Identifying High-Level Entities for Biological Metadata

Felicitas Löffler[1], Claas-Thido Pfaff[2], Naouel Karam[3], David Fichtmüller[4], and Friederike Klan[1]

[1] Heinz-Nixdorf Endowed Chair for Distributed Information Systems, FSU Jena, Germany
`{felicitas.loeffler,friederike.klan}@uni-jena.de`
[2] Systematic Botany and Functional Biodiversity Lab, University of Leipzig, Germany
`{claas-thido.pfaff@uni-leipzig.de}`
[3] Institute of Computer Science, Freie Universität Berlin, Germany
`{naouel.karam@fu-berlin.de}`
[4] Botanic Garden and Botanical Museum (BGBM), Freie Universität Berlin, Germany
`{d.fichtmueller@bgbm.de}`

Abstract. Research questions in biodiversity are as diverse and heterogeneous as data are. Most metadata standards are mainly data-focused and pay little attention to the search perspective. In this work, we introduce a method to analyze the actual information need of biodiversity scholars based on two individual studies: (1) a series of workshops with domain experts and (2) an analysis of research and search questions collected in three different biodiversity projects. We finally present 12 high-level entities that appear in all kinds of biological data across the different sources evaluated.

Keywords: biological data, life sciences, biodiversity, metadata, information retrieval

1 Introduction

In the last decade, we have witnessed an unprecedented increase of open data ranging from species-related observations, digitized specimen collections to genome or environmental data offered, e.g., through remote sensing. This opens up unforeseen opportunities particularly for biodiversity research, which relies on cross-disciplinary data analysis to elucidate the interplay between individuals and the conditions of the environment they inhabit, both on the macroscopic and microscopic level. At the flip side of this development, discovering and filtering these large volumes of multidisciplinary data becomes a more and more time-consuming and demanding task [2]. Thus, there are two big challenges: exploring effective retrieval mechanisms that support humans in finding relevant data and creating proper and rich metadata in order to make data findable (FAIR principles [6]).

The biodiversity community has responded to the latter requirement by developing metadata standards for biological data, such as Darwin Core (DwC)[5], ABCD[6] or EML[7] .

---

[5] Darwin Core, http://rs.tdwg.org/dwc/

[6] ABCD, http://www.tdwg.org/activities/abcd/

[7] EML, http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language

At the same time, considerable effort has been put on the formalization of domain knowledge in terms of vocabularies and ontologies. By referencing this formal knowledge, data can be richly annotated and become machine-readable. In the last years, numerous ontologies for specific biological domains have been created, e.g., the Gene Ontology (GO)[8] for genes, the Chemical Entities of Biological Interest (ChEBI) ontology for chemical compounds, the Environmental Ontology (ENVO) [9] for environmental features and materials, the Phenotype Quality Ontology (PATO) for phenotypes and the NCBI Taxonomy[10] for species. In addition, high-level ontologies with an emphasis on interlinking biological data from different sources have been developed, e.g., the Biological Collections Ontology (BCO) [5], the Extensible Observation Ontology (OBOE) [3] or the Semantic Sensor Network Ontology (SSN) [1].

In the search applications we are hosting within the biodiversity projects GFBio (The German Federation for Biological Data)[11] and AquaDiva[12], we observe that existing metadata standards and ontologies often take a data-centric view. They provide means to well-described biodiversity data, their characteristics, their origin and the process of their creation. However, when searching for data, scholars often do not have specific data in mind, but rather a research question they would like to answer. Hence, we argue that when designing metadata standards and ontologies for biodiversity both perspectives have to be considered, the requirements given by available datasets and the way scholars are looking for data. The contribution of the paper is twofold: (1) We propose and apply a method that combines the findings of two different and independent approaches to identify high-level entities that are relevant for biodiversity researchers when searching for data (Sects. 2.1 and 2.2). (2) As a first result, we present the findings of the two individual approaches and propose a consolidated set of biological entities (Sect. 2.3). We consider this as a first step towards enriched metadata with information that is relevant to information seekers. It also serves as a prerequisite for increasing the findability of biodiversity data.

2 Methodology

Our approach to analyze the search perspective comprises two independent studies: Assuming that properly described data can be found more easily, the goal in dedicated workshops with scholars was to define an annotation schema that can be used to richly describe ecological data (Sect. 2.1). The second study is oriented to evaluation methods in information retrieval and analyzes research and search questions collected in three biodiversity projects (Sect. 2.2). In both approaches, the aim was to enhance search applications and to detect high-level entities that can be either used as metadata fields or that can be linked with ontologies. The first result of biological high-level entities is presented in Sect. 2.3.

---

[8] GO, http://www.geneontology.org/

[9] ENVO, https://github.com/EnvironmentOntology/envo

[10] NCBI Taxonomy, https://www.ncbi.nlm.nih.gov/taxonomy

[11] GFBio, https://www.gfbio.org

[12] AquaDiva, http://www.aquadiva.uni-jena.de

2.1 Workshops with domain experts

In close collaboration between GFBio and the German Centre for Integrative Biodiversity Research Halle – Jena – Leipzig (iDiv)[13], we conducted ten workshops with 35 domain experts from ecology and adjacent disciplines to develop a metadata schema and a controlled vocabulary, the *Essential Annotation Schema for Ecology* (EASE)[14]. This annotation framework aims at describing ecological data from a scholar's search perspective. Annotation in this context refers to metadata.

Two design principles have been formulated for the development: *Parsimony:* The framework aims at being as simple as possible in structure and content. Optimization here has to be done carefully to maintain a differentiated and consistent annotation. One example: Larger time frames in ecology are referred to by a relative reference, (e.g., 18 million years ago) or by named geological time periods. These periods are getting more granular from eons to ages and are nested in each other. It could be argued to make ages optional in the annotation which sacrifices some granularity but still maintains a consistent larger temporal context. *Comprehensiveness:* The framework aims at a certain comprehensiveness defining essential orthogonal dimensions of information which allow ecological content to be described and located in the search space of ecology. Comprehensiveness is not accomplished by using many different dimensions and concepts but rather a few essential and complementary ones which also reflect the mindset and questions of researchers when looking for data.

Based on these guidelines, 8 top level categories have been selected. During the workshops, the top level categories were substantiated in a top down approach with increasing detail (~1600 concepts). Here, we relied on expert knowledge of the contributors but also on other sources such as EML, ABCD and DwC, various topic specific textbooks (e.g., related to organic and inorganic chemistry) and standardized vocabularies (e.g., the World Reference Base for Soil[15], and The International Chronostratigraphic Chart[16]).

The top level categories are 1. *Time* (e.g., date, time, timezone), 2. *Space* (e.g., bounding box, coordinates, location names), 3. *Sphere* (e.g., pedo-, hydro-, atmosphere aspects), 4. *Biome* (e.g., zones, water availability, land use), 5. *Organism* (species classification), 6. *Process* (e.g., processes, objects and interactions), 7. *Method* (general approach, setup of gradients), 8. *Chemical* (e.g., elements, compounds, functions). In addition, the framework covers a set of general information to handle associations between primary data and annotation (e.g., data format, contact person, download URL).

2.2 Research and search questions in the biodiversity domain

In information retrieval, a lot of research has been done towards a perfect ranking [4] whereas little attention has been paid to a user's actual information need. What research questions are biodiversity scholars working on? What kind of data do they want to reuse? Do the provided metadata actually reflect a researcher's information need? Therefore,

---

[13] iDiv, https://www.idiv.de/
[14] EASE: https://github.com/cpfaff/ease
[15] WRB, http://www.fao.org/soils-portal/soil-survey/soil-classification/world-reference-base/en/
[16] ICS, http://www.stratigraphy.org/index.php/ics-chart-timescale

Table 1: Example questions gathered in three biodiversity projects

gfbio	AquaDiva	iDiv
Do butterflies occur on calcareous grassland?	How does agriculture affect the groundwater composition?	How old does Plantago lanceolata get?
Is there data on the influence of geographic elevation on the growth rate and plant development of Zea mays?	What are suitable methods to characterize microbial soil processes by gas analytical techniques?	Do cities harbour a higher biodiversity compared to agricultural areas?

we collected 184 search and research questions from scholars who are involved in three biodiversity projects in Germany: GFBio (73), AquaDiva (98) and iDiv (13). Examples are presented in Table 1. We asked for full questions as well as keywords to get the actual information need together with the search query. We left it to the scholars to either provide search questions posed to a search interface or broader research questions they are currently working on to get a wider spectrum of information needs.

We analyzed the questions manually and explored whether the noun entities could be grouped into high-level categories, such as *Organism* or *Environment*. For instance, given the question: *Is there DNA data about Amphimonhystrella (Nematoda)?* the noun entities are 'DNA data' and 'Amphimonhystrella (Nematoda)'. The latter one is an *Organism* whereas 'DNA data' points to a certain *Data Type*. Finally, we grouped the noun entities into 13 categories presented in Table 2. *Organism* comprises all individual life forms including plants, fungi, bacteria, animals and microorganisms. All species live in certain *Environments* and have certain characteristics that are summarized with *Quality and Phenotype*. Biological, chemical and physical *Processes* are re-occurring and transform materials or organisms due to chemical reactions or other influencing factors. *Events* are processes that appear only once at a specific time, such as environmental disasters. Chemical compounds, rocks, sand and sediments can be grouped as *Materials and Substances*. *Anatomical Entities* comprise the structure of organisms, e.g., body or plant parts, organs, cells and genes. The term *Method* describe all operations and experiments that have to be conducted to lead to a certain result. Outcomes of research methods are delivered in *Data Types*. All kinds of geographic information is summarized with *Location* and time data including geological eras are described with *Time*. *Person and Organization* are either projects or authors of data. As reflected in the search questions, scholars in biodiversity are highly interested in *Human Intervention* on landscape and environment, e.g., fishery, agriculture.

2.3 Discussion

Table 3 constitutes a consolidation of the main entities identified by the previously described processes. While there is a broad consensus on entities such as *Organism, Process, Method and Time*, some wording and classification on others are different. *Space* in EASE actually means location information and *Sphere* comprises altitude indications that is covered under *Environment* in the search questions. In EASE, *Biome*

Table 2: Categories selected from search questions with examples below

Organism	Environment	Quality and Phenotype	Process	Event
quercus, cyclothone, globigerina bulloides	below 4000m, ground water, city	length, growth rate, reproduction rate, traits	climate change, nitrogen transformation	Deepwater Horizon oil spill, 'Tree of the Year 2016'
Material and Substance	**Anatomical Entity**	**Method**	**Data Type**	**Location**
sediment, rock, CO_2	DNA, proteome, root	lidar measurements, observation, remote sensing	lidar data, sequence data	Germany, Atlantic Ocean
Time	**Person and Organization**	**Human Intervention**		
current, over time, triassic	Deep Sea Drilling Project, author's or collector's names	agriculture, land use, crop yield increase		

contains subfields for *Land use* (*Human Intervention* in the search questions) and data attributes are defined as *Factor* under *Method* (*Quality and Phenotype* in the search questions). *Chemical* is grouped under *Material & Substance* in the search questions that additionally covers soil, sediments and rocks. *Anatomical Entity, Data Type* and *Event* only occur in the search questions.

Looking at potential linkage with existing ontologies, we finally selected 12 biological high-level entities. We left out *Sphere* since ontologies such as ENVO already cover environmental features and conditions and could be extended with altitude information. Preliminary, we will leave *Data Type* out. It needs to be further discussed and investigated whether it can be classified under other entities, e.g., *Method*.

3 Conclusion

We described and applied a methodology for identifying high-level entities in the biodiversity domain that reflect the scholars' point of view. In our future work, we will link the identified entities to existing ontologies. Our aim is to improve the indexing process of search applications over research data by means of these 12 categories. We would like to automatically extract information from metadata that is related to the entities. We believe, that this will help to improve data retrieval methods in the biodiversity domain.

Acknowledgements

Table 3: Consolidation of high-level entities

EASE	Questions	Consolidation
Organism	Organism	Organism
Process	Process	Process
✗	Event	Event
Environment	Environment	Environment
Method - Factor	Quality & Phenotype	Quality & Phenotype
✗	Anatomical Entity	Anatomical Entity
Chemical	Material & Substance	Material & Substance
Method	Method	Method
✗	Data Type	✗
Time	Time	Time
Space	Location	Location
Sphere	✗	✗
General Information	Person & Organization	Person & Organization
Biome - Land use	Human Intervention	Human Intervention

References

1. M. Compton, P. Barnaghi, L. Bermudez, R. Garca-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. L. Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor. The ssn ontology of the w3c semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:25 – 32, 2012.

2. M. Diepenbroek, F. Glöckner, P. Grobe, A. Güntsch, R. Huber, B. König-Ries, I. Kostadinov, J. Nieschulze, B. Seeger, R. Tolksdorf, and D. Triebel. Towards an integrated biodiversity and ecological research data management and archiving platform: GFBio. In *Informatik 2014*, 2014.

3. J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and V. F. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2007.

4. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

5. R. L. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, P. L. Buttigieg, N. Davies, D. Endresen, M. A. Gandolfo, R. Hanner, A. Janning, L. Krishtalka, A. Matsunaga, P. Midford, N. Morrison, E. O. Tuama, M. Schildhauer, B. Smith, B. J. Stucky, A. Thomer, J. Wieczorek, J. Whitacre, and J. Wooley. Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS ONE*, 9(3):e89606+, Mar. 2014.

6. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data 3*, (160018), 2016.

Toward better data sharing methods for genebanks

Evangelia Papoutsoglou[1], Rajaram Kaliyaperumal[2], Theo van Hintum[3],
Richard G.F. Visser[1], Ioannis N. Athanasiadis[4], Richard Finkers[1]

1 Plant Breeding, Wageningen University & Research, Wageningen 6708 PB, The Netherlands
2 Leiden University Medical Centre, Leiden 2333 ZA, The Netherlands
3 Centre for Genetic Resources, The Netherlands, Wageningen University & Research, Wageningen 6708 PB, The Netherlands
4 Information Technology Group, Wageningen University & Research, Wageningen 6706 KN, The Netherlands

Abstract. The conservation of plant genetic resources (PGR) is an important task that requires collaborative effort from many stakeholders. For this, common means of data exchange and effective methods to profit from the collected information need to be established. In this paper, we describe a demonstrator promoting findability of PGR, according to the FAIR (Findable, Accessible, Interoperable & Reusable) data principles. PGR providers can each expose their germplasm information, using the FAO Multicrop Passport Descriptor (MCPD), which subsequently can be queried in a distributed manner via a single user interface. PGR users can select among predefined questions, for example for specific crops, accessions or phenotypes.. On the back end, data integration from a distributed query is achieved through annotations with the MCPD semantics.

Keywords: MCPD, FAIR, germplasm, plant genetic resources, genebanks, interoperability, data modelling, metadata, linked data, semantic web

1   Introduction

Genetic diversity in crops, and the maintenance thereof, is a crucial factor for modern breeding research. However, access to information describing this genetic diversity is not always readily available. For example, many accessions can be obtained from genebanks worldwide. Each genebank has different means to document their accessions and how to make data available, which emphasizes the need to deal with this heterogeneity. The current solution includes documenting PGR data in aggregated systems, such as EURISCO and GENESYS, however, this is not a long-term sustainable solution as the volume of information is readily increasing, especially for (~omics-derived) characterization data. We believe that a way to gather and assemble data (smaller or bigger in size and/or complexity [1]) from distributed resources will be useful, and could significantly speed up the production of results in important genomic selection, genome-wide association studies and more [2]. So, in this paper, these challenges are addressed with a demonstrator interface, relying on the reuse of existing building blocks.

2 Background

To effectively work towards a better data sharing, two aspects need to be in place. The first is a data standard to effectively describe the data. For plant genetic resources (PGR), this is the multi crop passport descriptor (MCPD) vocabulary [3]. Secondly, we need a definition on what is required to promote optimal data management/stewardship, for as example defined in the FAIR data principles [4].

2.1 Findability of PGR passport data using the FAIR data principles

The FAIR data principles dictate that all data should be Findable, Accessible, Interoperable and Reusable. Findability is crucial for the discovery of information about PGR world-wide. and requires a well-defined data standard. For example, a PGR user might want to find accessions from a specific geographical region for a specified taxa. Within the PGR community, the MCPD vocabulary is the accepted community standard to describe PGR. The MCPD comprises a set of attributes describing an accession, such as accession identifier, taxon, geographical origin, holding institute, and biological status, uniformly describing PGR. In our work, we defined a FAIR data point definition (FDP), exposing PGR data with attached metadata in a semantic manner. We will show that exposing PGR passport data according to the MCPD standard utilizing the FAIR data principles will improve findability and subsequent querying of these resources, via a query interface targeted at PGR users. The application of the FAIR principles in a demonstrator is not novel. We reuse code from the FAIR rare diseases demonstrator [5] targeted at biobanking collections, where similar questions are raised (e.g. which biobank has samples from a patient having a certain disease phenotype). This approach also shows the added value of working with diverse communities on tackling common data challenges.

2.2 Use case-relevant plant semantics resources

1. Germplasm Ontology[1]: Contains parameters not included in the MCPD, e.g. distinguishing accessions or genotype, and describing these in more detail.
2. Agronomy Ontology[2]: Defining an experiment, as a container to bind together material with other parameters (e.g. treatments, environment, etc.)
3. MIAPPE[3] (and its implementation, the Breeding Application Programming Interface - BrAPI): further describes the organization of an experiment, and holds more attributes describing it.
4. Plant Ontology[4]: generic ontology for plant structure and anatomy.
5. Plant Environment Ontology[5]: for treatments and growing conditions in plant biology experiments

---

1   http://www.cropontology.org/ontology/CO_010/Germplasm

2   http://www.obofoundry.org/ontology/agro.html

3   http://www.miappe.org/

4   http://purl.bioontology.org/ontology/PO

5   http://www.obofoundry.org/ontology/eo.html

6. Plant Stress Ontology[6]: for diseases and pathogens
7. Plant Trait Ontology[7]: generic ontology for the description of phenotypic traits in plants, with mappings to crop-specific trait ontologies.
8. Other species-specific ontologies, where the Trait Ontology may prove insufficient. For example, in the case of tomato, an ontology like the Solanaceae Phenotype Ontology[8] may be used for more crop-specific attributes and more agile development from that specific community.

3 Methodology and Results

3.1 Model building and choices

The main challenge was to design a model incorporating the MCPD, and attached characterization data from a (field) experiment. To do this, we identified the ontologies listed above.

Table 1. Model in terms of triples. Terms starting with a colon (:) are instances of a class, quotes ("") enclose literal values, and brackets (<>) are used to refer to classes. Italics indicate placeholder terms, modeled specifically for this application.

#	Subject	Predicate	Object
1	:experiment_X	rdf:type	<AGRO:agricultural_experiment>
2		geo:long	"longitude"
3		geo:lat	"latitude"
4		dct:identifier	"ID"
5		dct:created	"creation date"
6		RO:has_participant	:plant_X
7		SIO:is_source_of	:observation_
8	:observation_X	rdf:type	<om:observation>
9		SIO:is_about	:plant_X
10		to:has_phenotype_score	"value"
11		to:has_phenotype_variable	:trait_X
12	:trait_X	dct:title	"trait_title"
13	:plant_X	rdf:type	*<plant>*
14		dct:identifier	"plant_ID"
15		*descendant_of*	:plant_Y
16	:plant_Y	*has_biological_status*	<MCPD_status>
17		*has_id*	:plant_identifier
18		*has_genus*	"genus"
19		*has_species*	"species"
20		*has_taxon_id*	<NCBI_ID>
21	:plant_identifier	rdf:type	<Accession (germplasm ontology)>
22		dct:identifier	"accession_ID"
23		*is_stored_in*	*<database>*

---

6   http://wiki.plantontology.org/index.php/Plant_Stress_Ontology

7   https://bioportal.bioontology.org/ontologies/PTO

8   http://www.cropontology.org/ontology/SP/Solanaceae%20Phenotype%20Ontology

Model structure. The developed model is presented in terms of semantic triples. The passport information is given with the MCPD, and AGRO is indicated for the experiment. Ontologies for the domain of plant breeding (Trait Ontology, Plant Ontology) can be used across crops, and supplemented by other crop-specific ontologies (Solanaceae Phenotype Ontology). Widely-used ontologies (prefixes rdf, geo, dct – dublin core terms, RO - Relations Ontology, SIO - Semantic science Integrated Ontology) could be used for generic terms. However, many predicates appropriate for this use, have not been defined in published ontologies, hence the need for placeholders.

The core of this model is the Experiment. For the sample queries presented in section (3.2) this is not necessary, but it allows the model to be easily extended with, for example, treatments and management. For now, the date of the experiment, its location and title are attached to it. Each experiment has a set of observations, each of which is made on a specific, physical plant or accession. The observation consists of one phenotypic variable, and the value for the trait being observed. Each plant has a local identifier for the specific experiment. It may possess more broadly used identifiers, through its link to, for example, an accession. To cover the cases of crosses, an ancestral plant is introduced, as a descendant of the physical entity in the experiment. In the case where an accession identifier is not used, another property (like genotype) might be used instead. Further MCPD attributes (such as common crop name, institute information, addresses and coordinates, taxon authorities etc. would also be specified here, including the holding institute from which this identifier originated.

Placeholders. Many terms in the table do not refer to a specific ontology or vocabulary. Especially for predicates, the lack of suitable terms is a hindrance for good semantics. Even in the prominently featured MCPD, those are lacking, and do not give any means (properties) to connect an entity with, for example, its biological status - though they do contain the relevant classes. Issues like this may already be subject of attention, but have not yet been resolved. Additionally, it is noted that these terms come from a variety of ontologies, the terms of which are not defined to be compatible. Therefore, it is imperative that, for such an example to be semantically correct, this needs to be amended, and constraints need to be more appropriately defined.

3.2 The demonstrator

The demonstrator itself, (Fig. 1) uses the above semantic model, and data available from the EU-SOL database (https://www.eu-sol.wur.nl/). The example questions were formulated in collaboration with PGR users and PGR providers, and they all require to query the MCPD for the "accessions" and the "crops", as well as location data. These questions were hard-coded in the demonstrator (Fig. 1). As a user, one has to select the relevant query and specify its parameters (like the phenotype to search for, the desired country of origin, biological status, accession name). As we focused on tomato, the Solanaceae Phenotype Ontology was used. The options for each parameter are queried on the fly and displayed in a drop-down list. Accordingly, a SPARQL query is formulated, and run against the provided sources.

Limitations. The demonstrator currently does not search for relevant datasets across FAIR data points by itself. Instead, it retrieves the data from hard-coded resources, in the form of RDF, formatted according to the FDP definition. However, the demonstrator will be adapted to consume data from distributed resources once the relevant FDP's, formatted according the heretofore mentioned data model, are coming online; which also would enhance the possibility to query these resources directly by machines (e.g. via the SPARQL query language). The demonstrator is online at https://www.plantbreeding.wur.nl/ld-demonstrator/.

Fig. 1. The demonstrator interface: the user selects a question (highlighted), a phenotypic variable ("fruit color"), as well as a value for it ("red")

4   Discussion

Outcome. This demonstrator was developed to showcase how FAIR data infrastructures contribute to the sharing of PGR data. The result is a responsive graphical interface, answering predetermined questions but allowing more flexible querying via SPARQL queries directly. The value of this effort does not come from any novel questions posed, but from the distributed nature of the available PGR resources. Work on the semantic data model brought up some significant gaps that currently exist in the semantics that should be addressed in the future, such as the

placeholders in Table 1. In spite of those, the approach followed is a good example of such a process, highlighting the reusability of existing components.

Modeling pitfalls. The most demanding part is the construction of a semantic model. Lessons learned include: one should not deviate from designing a model reflecting the "real world" conditions, in favor of modelling for a specific dataset or database. This is to reaffirm that a specific database or entity-relationship diagram (ERD) is easily translatable into semantic triples, but does not necessarily lend itself to a schema that is intended to accommodate data from different providers.

Future work. In the future, the demonstrator will be extended to include more domain-relevant queries and implementation of the FDP infrastructure by PGR providers. As plants are "unable to move", we plan to explore the potential of geo-aware queries. However, the main challenge will be in the full integration with other data sources, such as weather or especially ~omics databases. Only then could big data technologies help to revolutionize plant breeding and have a significant impact on the world's food and nutrient security.

References

1. Gray, E., Jennings, W., Farrall, S. & Hay, C. Small Big Data: Using multiple data-sets to explore unfolding social and economic change. Big Data & Society January-Ju, 1–6 (2015).
2. Spindel, J. E. & McCouch, S. R. When more is better: how data sharing would accelerate genomic selection of crop plants. New Phytologist (2016). doi:10.1111/nph.14174
3. FAO/Bioversity, Multi-crop passport descriptors v.2. (2012) [Online], Available: http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/, [Accessed: 8-Sept-2017].
4. The Editors. FAIR principles for data stewardship. Nature Genetics 48, 343–343 (2016).
5. Roos, M., Wilkinson, M., Kaliyaperumal, R., Thompson, M., Carta, C., Cornet, R. and da Silva Santos, L.O.B., Registries of domain-relevant semantic reference models help bootstrap interoperability in domains with fragmented data resources, Proceedings of the 9th International Conference Semantic Web Applications and Tools for Life Sciences, Amsterdam, The Netherlands, December 5-8, 2016.
6. Khoury, C., Laliberté, B. & Guarino, L., Genetic Resources and Crop Evolution (2010), Volume 57, Issue 4, pp 625–639 https://doi.org/10.1007/s10722-010-9534-z
7. Mackay M. C. (2011), Surfing the Genepool: the Effective and Efficient Use of Plant Genetic Resources, Doctoral thesis No. 2011:90, Acta Universitatis Agriculturae Sueciae, Swedish University of Agricultural Sciences, Faculty of Landscape Planning, Horticulture and Agricultural Science, Alnarp, Sweden.

EcoPortal: a proposition for a semantic repository dedicated to ecology and biodiversity

Nicola Fiore[1] [0000-0002-9538-2966], Barbara Magagna[2] [0000-0003-2195-3997] and Doron Gold-farb[2] [0000-0003-1183-6041]

[1] LifeWatch Italy, University of Salento, Lecce, Italy
[2] Umweltbundesamt GmbH, Vienna, Austria
nicola.fiore@unisalento.it, barbara.magagna@umweltbundesamt.at,
doron.goldfarb@umweltbundesamt.at

Abstract. This paper presents the joint effort of LifeWatch Italy and LTER-Europe to design EcoPortal, a semantic repository focused on ecology and biodiversity as well as on ecosystem observation mainly in the European context. It is our aim to offer a space to collect domain ontologies as well as thesauri and domain relevant reference lists. We plan to test NCBO BioPortal technology to accommodate community requested functionalities.

Keywords: Registry · Ontology · Thesaurus ·Reference List · Semantics.

1 Introduction

To address today's ecological challenges, it is necessary to use data coming from different disciplines and providers. Thus, discovery and integration of data, especially from the ecological domain, is highly labour-intensive and often ambiguous in semantic terms. To improve the discovery, integration and re-usability of data the use of semantic resources can help to harmonise and enrich the description of datasets and its content. In the last decade research groups and infrastructures focusing in the monitoring and analysis of ecosystem properties have increasingly put effort into the development of semantic resources mainly based on core ontologies such as OBOE or the O&M data model [1].

This paper presents the joint intention of the European networks LifeWatch Italy[1] and LTER-Europe [2] to design a vocabulary repository focused on the ecology and biodiversity research as well as on observation of biological and physical-environmental data. This initiative will support the community in the management and integration/alignment of their semantics and subsequently also of their data [3].

In order to increase interoperability between different domains and institutions, LifeWatch Italy and LTER-Europe[2] developed ontologies (LifeWatch Ontology[3] and

---

[1]   http://www.lifewatchitaly.eu
[2]   http://www.lter-europe.net/
[3]   http://semantic.lifewatchitaly.eu

SERONTO[4]) as a semantic framework for integration of monitoring [4] and biodiversity data and common vocabularies for harmonised data annotation (LifeWatch-Italy Thesauri[5], concerning functional traits, and EnvThes[6] - Environmental Thesaurus [5]).

LifeWatch Italy and LTER-Europe are collaborating in order to improve and extend the existing thesauri and trace the semantic relations between them. In this context, the lack of a common semantic repository for the ecological domain became evident. We envisage to build a semantic platform for the domain to support not only the joint work done by the infrastructures, but to be a robust and stable reference repository for the European ecological community.

2 State of the art

Scientific communities are using an increasing number of ontologies or controlled vocabularies to disambiguate the description of data. To make these vocabularies discoverable and usable by software or by the community, different approaches exist:

- Distributed RDF stores with SPARQL endpoints allowing to access vocabularies using SPARQL queries. The presence of such endpoints does not solve the issue of discoverability, as you must already be aware of the semantic resources by other means to make use of them. To a certain extent they facilitate interoperability between different semantic resources but this requires high familiarity with the data representation schema and the granularity of each federated source.
- Semantic repositories (known also as ontology libraries [6]) are centralised access points providing both discoverability and access to semantic resources.

The latter, on which this paper focuses on, are collections of ontologies and thesauri with the primary purpose of enabling users to find and use them. They should be distinguished from ontology search engines, such as Swoogle[7], which automatically crawl the Web to index ontologies rather than collect them. Also, we want to exclude here collections on data, such as the Linked Open Data collection of datasets[8].

According to the targeted user set approach of d'Aquin and Noy [6] different types of repositories can be identified, although they often exist in a mixed form: curated directories, registries and application platforms.

Many repositories already offer additional services, the most prominent ones are BioPortal[9] and EBI OLS[10].

---

[4]http://www.umweltbundesamt.at/fileadmin/site/daten/Ontologien/SERONTO/SERONTOCore20090205.owl

[5] thesauri.lifewatchitaly.eu/

[6] http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn

[7] http://swoogle.umbc.edu/2006/

[8] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

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

[10] https://www.ebi.ac.uk/ols/index

We would also like to emphasise that most of the repositories are dedicated solely to ontologies, some only contain thesauri like Finto[11] and LusTRE[12] and only a few seem to offer the place to publish both ontologies and thesauri like AgroPortal[13]. The inclusion of thesauri is important in our considerations because they are essential sources of harmonised knowledge (not only) in the ecology domain.

3   EcoPortal

3.1   Requirement Elicitation: Purpose and Coverage

The main goal of the EcoPortal initiative is to provide a central registry for semantic resources (e.g. vocabularies) used in the ecological and biodiversity domain allowing users to identify and select semantic resources for specific tasks, as well as offering generic services to exploit them in search, annotation or other scientific data management processes.

To reach this objective the user-centred, structured and systematic approach AWARE (Analysis of WebApplication Requirements) has been adopted [7].

Following the AWARE guidelines, the following main stakeholders (i.e. user profiles to be considered for the Web application) have been identified:

— *Domain Expert*, is the user of the portal and expert of the ecological domain. One high-level goal of this kind of user is to explore the semantic world in the ecological domain to understand how to annotate experimental data to enable interpretation, comparison, and discovery across databases. For this kind of user it is necessary to offer very user-friendly tools and services.
— *Semantic Author*, is a domain expert user that creates and shares a specific vocabulary/ontology and is responsible to maintain it updated.
— *Semantic Engineer*, is a type of user with semantic technology skills, who aims to design new tools/services for the domain expert.
— *System Owner,* who creates and manages EcoPortal and its services.

Figure 1 shows parts of the requirements analysis made for the stakeholder *Domain Expert*. For each stakeholder we have identified goals and tasks (i.e. high-level user activities on the site) and in the refined process they have been recompiled into requirements. We can classify and synthesise the main requirements of the EcoPortal in the following categories.

- *Content Requirements*
  The focus of the portal will be on the ecology, ecosystem and biodiversity domains. Not only ontologies but also thesauri will be collected and managed. Each semantic resource will be described by metadata (i.e. Structure Content

---

Requirements in AWARE). The need of a common metadata set has been identified by several initiatives like OBO Foundry [8], LOV[14] and AgroPortal.

- *Access Path and Navigation Requirements*
  - Different search paths should be supported fitting the general requirements: search within and across ontologies/thesauri, structured search via a SPARQL query engine and advanced search will be developed. In the scenario in Figure 1 the *Domain Expert* needs to perform a search for "equivalent terms" and to navigate from a term to the related one.
  - To facilitate the semantic resource discoverability, we want to use categories (ecology, observation, etc.) as used in AgroPortal.
  - Browsing functionality will be offered including different types of visualisation of the content. So, we aim to foresee automatic translations for single terms wherever the vocabularies provide multilingual labels for them. Access services will be also provided for all the resources, including ontologies/thesauri metadata and the mappings between them.
  - We also intend to collect reference lists codified in SKOS used to define permissible values in certain data fields, providing information needed to make other data meaningful and interpretable in an unambiguous way. Translating from one reference list to another within the same domain is an essential need for ecologists.
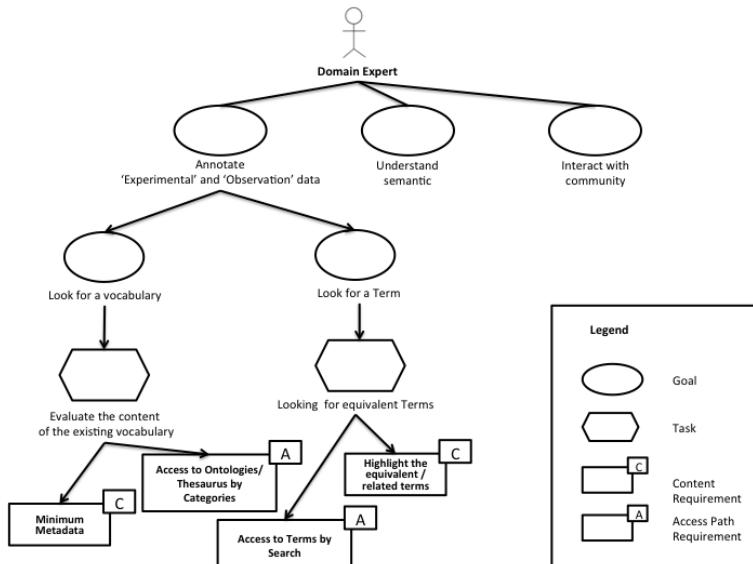

Fig. 1. Requirements for Domain Expert

- *System and User Operation Requirements*
  - The portal should enable automatic (based on exact matching labels) as well as manual mappings between semantic resources (private and/or public accessible -

storing also metadata on mappings), and it should allow upload of mappings created elsewhere.

- For collecting resources, EcoPortal should use a hybrid approach: apart from the administrators (ensuring to host the newest version of the resource also published in other portals) also users should be enabled to submit their resources to the collection through a dedicated user interface.
- As far as belongs to gatekeeping, we envisage a two-step approach: after uploading, the semantic resource is validated by a quality committee, after that it is published in the catalogue. But before validation the resource should already be visible to the users labelled as not yet validated. Quality requirements should include metadata description, syntactically correctness and thematic relevance.
- The portal should be able to automatically compute ontology metrics.
- It should enable social interaction, allowing comments on ontologies and components (at class level).
- Instead of ranking ontologies by their relevance, we would prefer an exchange information platform between supplier and user where it should become clear for which use cases the resources were originally developed and then used. This concept of semantic marketplace has been introduced at the EUDAT Semantic Workshop[15]. We want to encourage developers to publish their vocabularies in our Portal in an early stage of their development taking advantage of the domain community.

3.2   Expected Contents.

A first inventory of the appropriate and relevant ontologies, thesauri and reference lists to be hosted in the repository can be accessed online[16]. This list will be extended by community contributions as collaborative and open process.

3.3   Conclusions and Future Work

The paper briefly introduces the ongoing work of LifeWatch Italy and LTER-Europe in order to develop EcoPortal, a semantic repository focused on the ecosystem and biodiversity research as well as on observation of the ecosystem. A common domain specific repository of semantic resources allows their better integration into the workflows of metadata annotation (e.g. DEIMS-SDR[17]) and discovery. This fosters the semantic interoperability not only on the metadata but also on the data level.

---

[15] https://www.eudat.eu/events/trainings/co-located-eudat-semantic-working-group-workshop-9th-rda-plenary-barcelona-3-4

[16] http://www.servicecentrelifewatch.eu/web/ecoportal/wiki/-/wiki/Main/EcoPortal+semantic+resources?_36_redirect=http%3A%2F%2Fwww.servicece ntrelifewatch.eu%2Fweb%2Fecoportal%2Fwiki%2F-%2Fwiki%2FMain%2Fall_pages%3Fp_r_p_185834411_title%3DEcoPortal%2Bsemantic%2Bresources

[17] https://data.lter-europe.net/deims/

A first prototype in line with the described architecture is planned to be online by October 2017. In the initial phase, we will test the NCBO BioPortal technology to accommodate community-requested functionalities with semantic resources of European networks. Considering the importance of such tools in the ecological field, we expect a broad adoption of the EcoPortal in the community in the long run. Furthermore, LifeWatch as ERIC will be able to assure the long-term product sustainability.

The most pressing issues still to be addressed are the ability to manage and search across different types of semantic resources like OWL ontologies and SKOS thesauri as well as the use of a minimal metadata set and of a vocabulary marketplace considering the ongoing discussions in the RDA VSIG[18].

4   Acknowledgments

References

1. Cox, S.J.: Ontology for observations and sampling features, with alignment to existing models. Semantic Web, 8(3), 453-470 (2017).
2. Mirtl M.: Introducing the Next Generation of Ecosystem Research in Europe: LTER-Europe's Multi-Functional and Multi-Scale Approach. In: Müller F., Baessler C., Schubert H., Klotz S. (eds) Long-Term Ecological Research. Springer, Dordrecht (2010). doi: 10.1007/978-90-481-8782-9_6
3. Oggioni, A., Carrara, P., Kliment, T., Peterseil, J. & Schentz, H.: Monitoring of Environmental Status through Long Term Series: Data Management System in the EnvEurope Project. In: Proceedings EnviroInfo 2012, pp. 293-301. Sahker Verlag, Aachen (2012).
4. Van der Werf, B., Adamescu, M., Ayromlou, M., Bertrand, N., Borovec, J., Boussard, H., et al.: SERONTO: A Socio-Ecological Research and observation oNTOlogy. In: Weitzman, A. L. & Belbin L. (eds.) Proceedings of TDWG. Fremantle, Australia (2008).
5. Schentz, H., Peterseil, J., Bertrand, N.: EnvThes - interlinked thesaurus for long term ecological research, monitoring, and experiments. In: Proceedings EnviroInfo 2013: Environmental Informatics and Renewable Energies, pp. 824-832. Shaker Verlag, Aachen (2013).
6. d'Aquin, M., Noy, N.: Where to Publish and Find Ontologies? A Survey of Ontology Libraries. In Web Semant 11: pp. 96–111 (2012). doi: 10.1016/j.websem.2011.08.005.
7. Bolchini, D., Paolini, P.: Goal-driven requirements analysis for hypermedia-intensive Web applications. In: Requirements Eng 9: pp. 85–103 (2004). doi: 10.1007/s00766-004-0188-2
8. Smith, B., Ashburner, M., Rosse, et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. In: Nature Biotechnology, 25 (11), 1251 (2007).

---

[18] https://www.rd-alliance.org/groups/vocabulary-services-interest-group.html