Datenbank Spektrum (2016) 16:195–205
DOI 10.1007/s13222-016-0231-8

SCHWERPUNKTBEITRAG

A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data

Naouel Karam[1] · Claudia Müller-Birn[1] · Maren Gleisberg[2] · David Fichtmüller[2] · Robert Tolksdorf[1] · Anton Güntsch[2]

Received: 30 June 2016 / Accepted: 19 September 2016 / Published online: 5 October 2016

Abstract Research has become more data-intensive over the last few decades. Sharing research data is often a challenge, especially for interdisciplinary collaborative projects. One primary goal of a research infrastructure for data management should be to enable efficient data discovery and integration of heterogeneous data. In order to enable such interoperability, a lot of effort has been undertaken by scientists to develop standards and characterize their domain knowledge in the form of taxonomies and formal ontologies. However, these knowledge models are often disconnected and distributed. The work presented here provides a promising approach for integrating and harmonizing terminological resources to serve as a backbone for a platform. The component developed, called the GFBio Terminology Service, acts as a semantic platform for access, development and reasoning over internally and externally maintained terminological resources within the biological and environmental domain. We highlight the utility of the Terminology Service by practical use cases of semantically enhanced components. We show how the Terminology Service enables applications to add meaning to their data by giving access to the knowledge that can be derived from the terminologies and data annotated by them.

✉ Naouel Karam
naouel.karam@fu-berlin.de

✉ Maren Gleisberg
m.gleisberg@bgbm.org

[1] Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany

[2] Botanic Garden and Botanical Museum Berlin, Freie Universität Berlin, Berlin, Germany

Keywords Research data infrastructure · Interoperability · Data discovery · Terminology repository · Ontology-based data access · Ontologies · Taxonomies

1 Introduction

Research practice has become increasingly more data-intensive over the last few decades and this methodological change is often referred to as the "fourth paradigm" for scientific exploration [16]. It describes a research practice in which data are collected, for example, by instruments and then processed by software. The resulting data are stored temporarily or in long-term archives. Scientists finally analyze these data in order to find meaning in them. This process is rarely carried out by a single researcher; science is more of a collaborative endeavor and scientists construct knowledge collaboratively. The foundation for this collaborative knowledge construction is the data.

However, these data might either be unavailable and expensive to capture or available but distributed across several archives. In the latter case, researchers need to know about potential data providers and need to develop an understanding of the data structures and content. For example, the availability for ecosystem science of high spatial resolution remotely sensed data is also changing environmental science [16]. Furthermore, data made available in one discipline allows researchers in another discipline to ask completely new questions and derive novel insights. At the same time, research questions are becoming more complex, and this requires researchers to collaborate despite geographical distribution and disciplinary boundaries.

In order to address these challenges, Hey and colleagues [16] call for software tools that cover the aforementioned activities from capture and data validation to analysis and

permanent archiving. These tools need to build upon a new research infrastructure. Such infrastructure would allow researchers to focus on their scientific questions rather than on the data management process [3]. The sustainability of such infrastructures and the durable integration into existing research practices is still a challenge today.

In our paper, we address this challenge in the biological and environmental research area that is distinguished by a large number of disciplines (e.g. Botany, Zoology, Microbiology, Chemistry, Geo-Sciences), dealing with highly heterogeneous and disparate data sets. Significant amounts of data have been created at all scales of research, from an individual researcher to an international research program.

These data are partly available in data archives that can be small institutional data repositories or globally acting, certified data centers. However, it is not only this fragmentation of existing data archives that makes data acquisition challenging for researchers. At least as important are the different scientific terms that are used within the different disciplines for the same concept or the same name referring to different concepts [6].

The German Federation for Biological Data (GFBio) was founded[1] in the light of these challenges. GFBio aims at providing data management and data archiving solutions for data capture, annotation, indexing, searching and storage. These solutions range from tailored Excel spreadsheets to virtual research environments, such as the Diversity Workbench [33], the Bexis system [14] or the EDIT Platform for Cybertaxonomy [7].

In this paper, we present our vision of a semantically enriched data management and archiving solution for GFBio by introducing a semantic-aware research infrastructure adopting a design science approach [15]. First, we describe the general concept of a semantic-aware research infrastructure, and then we derive a set of requirements regarding semantic services that can support such an infrastructure. We review existing systems in the light of those requirements, then the general architecture of our proposed solution is presented together with implemented use cases and their ongoing evaluation. Finally, we discuss and conclude by highlighting the added value of using a semantic-aware solution in the context of our project.

2 Towards a semantic-aware research infrastructure

In this section, we envision a semantic-aware research infrastructure that augments the traditional approach of organizing data by semantic technologies. Firstly, we highlight the general architecture of this infrastructure by describing the GFBio project and, secondly, we derive requirements for the semantic enhancement of this infrastructure.

2.1 German Federation for Biological Data (GFBio)

GFBio [10] is developing an infrastructure to enable biological and environmental scientists to share and discover their data more efficiently. In our vision, this infrastructure is being extended by semantic components that ensure, in addition to efficient data capture and discovery, the interoperability of data that are extremely heterogeneous in their structure, formats and meaning. Figure 1 presents an overview of the *semantic-aware research infrastructure* of GFBio, consisting of four main components.

The *GFBio Repository Network* (upper right in Fig. 1) comprises amongst others molecular data (EMBL-EBI[2]), environmental data (PANGAEA[3]), as well as natural history and culture collection data (e.g. MfN[4], DSMZ[5] and SNSB[6]).[7]

These data repositories register their data through the *GFBio Data Portal* (upper right in Fig. 1). GFBio builds upon the existing data infrastructures and augments the capabilities of the data archives by connecting them. One advantage of this registration approach is that existing data management practices do not need to change; they can evolve smoothly.

The GFBio Portal provides researchers with services such as *indexing*, *annotating* and *searching* data sets. These services help researchers upload, publish, share and discover their data in an efficient way. The data provided are indexed and semantically enriched, which allows for a global and efficient access of those independent from their original context. For researchers, this approach provides a "meaning" for the data. Understanding the meaning of the data allows scientists to integrate, analyze and visualize them, for example, by using the *GFBio VAT System* (upper left in Fig. 1) [4].

All these components are based on the assumption that the "meaning" of the data is provided by a fourth component – the GFBio Terminology Service (bottom left in Fig. 1). This service has been built upon considerable efforts that have been undertaken by scientists to describe their domain knowledge in well-structured ontologies and

[1] www.gfbio.org

[2] The European Bioinformatics Institute (www.ebi.ac.uk)

[3] Data Publisher for Earth & Environmental Science (www.pangaea.de)

[4] Natural History Museum (www.naturkundemuseum.berlin)

[5] German Collection of Microorganisms and Cell Cultures (www.dsmz.de)

[6] The Bavarian Natural History Collections (www.snsb.mwn.de)

[7] The complete list of involved archives and data centers is available on the GFBio website.

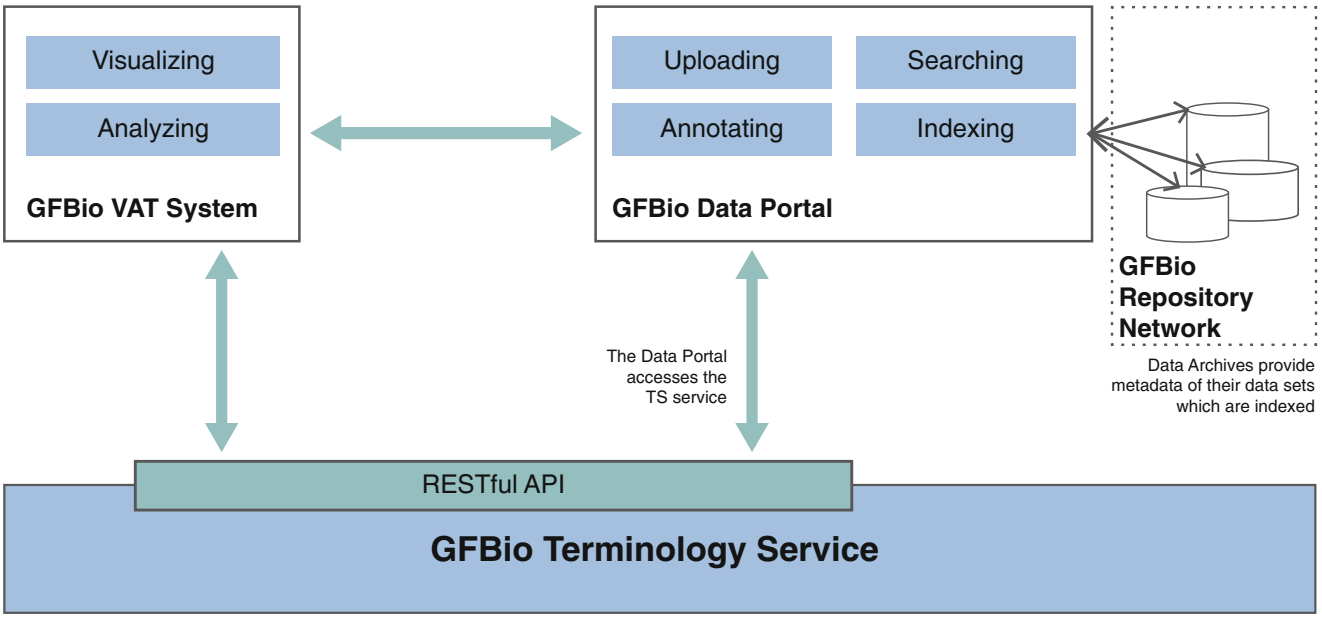

Fig. 1 Semantically enriched components of the GFBio Data Portal

taxonomies. Ontologies are formal specifications conceptualizing a shared area of interest. They have become a fundamental resource in the biological and biomedical domains, where they provide a framework for data annotation, discovery and curation. Another valuable resource in biological studies and biodiversity management are taxonomies of biological organisms as a basis for classification. However, these ontologies and taxonomies are created and used in specific domains of expertise. A semantic-aware research infrastructure needs to create the necessary links across such vocabularies (ontologies and taxonomies) by linking information as part of a global network of facts.

Taxonomies, for instance, have been combined with ontologies to achieve a taxonomy-based partitioning of the Gene Ontology (GO) [23] or formalize taxon-based constraints to detect inconsistencies and improve the quality of ontologies and annotations [8].

The interoperable exchange of information is vital to share knowledge between scientists more successfully. This provision and interlinking of ontologies and taxonomies is enabled by the GFBio Terminology Service (GFBio TS), the semantic component in the GFBio infrastructure that cross-links existing vocabularies and provides "meaning" to heterogeneous data.

In the following, we introduce a number of basic design requirements for the GFBio TS, and then compare these requirements to existing terminology repositories and services in the subsequent section.

2.2 Requirements on a Terminology Service

In the GFBio infrastructure, the TS constitutes the terminological and ontological basis for data annotation, discovery and curation. We defined a set of basic requirements for the TS to archive our vision of a semantic-aware research infrastructure that follows a semantic-aware approach in the context of GFBio. These requirements are described next.

Provide a single access point to heterogeneous terminological resources Scientists in the biological and environmental domain make use of heterogeneous vocabularies, such as formal ontologies and taxonomies, various concept collections or so-called informal ontologies (such as locations available via Geonames). These are available in a range of different formats with varying degrees of semantic interoperability. In order to allow researchers to share, analyze and access data sets provided by their peers easily, the GFBio TS needs to integrate all these vocabularies and provide a single access point to them. The service should deliver the information requested from the various terminological resources in a unified form.

Allow for query expansion and semantic data access When searching for existing data sets, scientists may use different terms to refer to the same concept; it is often challenging to find the correct query terms in order to discover relevant data sets. The GFBio TS should allow accessing data even if the terms annotated differ from the ones in a user's query or when the same term is used to refer to different concepts. Such data access can be performed either by query expansion, which can be sufficient in some

application settings, or via an ontology-based data access, which is a structured query enrichment with information that can be derived from one or more ontologies.

Enable efficient semantic annotation Researchers should be supported in their data publication efforts to lower the barriers for participation. In order to facilitate data integration, data needs to be annotated based on existing terminologies. Hence, researchers need an excellent understanding of the existing terminologies, their coverage, their added value for data discovery and data presentation, and their quality and development status. The GFBio TS should support scientists by delivering the optimal term for annotation based on the criteria cited above; in addition to term names, it should also cover synonyms, common names and abbreviations.

Enrich project-related terminologies Scientists create their own terminologies based on their project needs and their research context, especially in small-scale research projects. In these cases, the integration of the data into a well-defined terminological environment is often challenging if not impossible. The GFBio TS should offer a set of tools for supporting the development, curation and publication of such terminologies. This set of tools include, for instance, transformation tools from textual and tabular documents into a semantic format, a linked data interface, terminology integrity checks and validation, etc.

Provide mappings between terminologies Since scientists use their own vocabulary when describing their data sets, they may also use different terminologies for annotating them. It is crucial for an efficient discovery of such annotated data that the terminologies used in the annotation are interlinked. The alignment of terminologies is a key task when dealing with highly heterogeneous data sets [11]. It is defined as a process of determining correspondences between terms from different terminologies. Those alignments serve as a basis for efficient data integration and access.

Access to reasoning capabilities Annotated data can be more efficiently accessed in specific application scenarios using the logical reasoning capabilities enabled by ontologies. Indeed, additional information can be derived from ontologies statements allowing the extension of the search capabilities way beyond a restricted set of keywords, enabling a more valuable and precise data discovery for scientists. The GFBio TS should give access to some knowledge that can be derived from its underlying ontologies via reasoning, particularly for ontologies that fall under the OWL 2 EL profile for which reasoning can be performed in polynomial time with respect to the size of the ontology [5].

3 Existing terminology repositories and interfaces

Based on the requirements described in the previous section, we looked for and examined existing systems providing a comparable terminology service. These systems can be either full platforms for terminology management or frameworks for accessing them. We summarize our findings in Tab. 1. Each row in the table corresponds to one of the requirements, except for rows two and three, since both relate to the second requirement. Each one of the seven columns shows to what extent the respective system fulfills the respective requirement. A system meets the requirement either fully (filled circle), partly (crossed circles) or not at all (empty cell). Next, each of the systems is briefly introduced and its functionality is discussed in the light of the requirements defined.

Bioportal[8] [27] is a widely used ontology repository that provides browsing capabilities to a large number of biomedical ontologies and also a set of web services for accessing them. Bioportal is designed to store multiple versions of an ontology and it offers access to historical versions. It enables the community to participate in the evaluation and evolution of the ontologies provided: For instance, it offers term mapping capabilities, commenting tools and ontology reviews. Additionally, the annotator web service processes texts automatically by annotating strings in a text with terms from Bioportal ontologies.

Finto[9] (Finnish thesaurus and ontology service) [31] is a vocabulary service that resulted from the deployment of the ONKI Ontology service [35] into a sustainable national service. Finto/ONKI offers a set of interfaces and services for the publication and utilization of vocabularies, ontologies and classifications. Terminologies maintained by this system pertain to different domains, such as art, geography, science and medicine. Furthermore, general ontologies have been developed in order to enable interoperability between the various domain ontologies.

The Ontology Lookup Service (OLS)[10] [9] is a cross-platform system integrating publicly available biomedical ontologies in a single database. It can be accessed interactively via a web-based user interface or programmatically via a set of web services. Ontologies are loaded into a local database on a daily basis. Relevant information is extracted from the original files and persisted locally; this includes term names, synonyms and relationships with other terms.

[8] bioportal.bioontology.org

[9] www.finto.fi

[10] www.ebi.ac.uk/ols

Table 1 Existing terminology repositories & interfaces requirements coverage (● available, ◐ partly available)

	Bioportal	Finto/ ONKI	OLS	Ontobee	Aber-OWL	NOR	OntoCat
Heterogeneous Terminologies						●	
Query expansion services	●	●	●	●	●		●
Semantic data access					●		
Semantic annotation	◐	◐	◐	◐	◐	◐	◐
Terminologies management	●	●	●	●			
Terminologies mapping	●	●					
Reasoning capabilities							

The core functionality of the OLS is to enable users to perform queries using its underlying ontologies and to navigate relationships between concepts.

Ontobee[11] [37] is a linked data server and browser for ontology concepts. It provides different output formats and an HTML output of concept details and their RDF/XML representation and, thus, makes them accessible for Semantic Web applications.

Aber-OWL[12] [18] is a framework that provides reasoning services over bio-ontologies. It consists of an ontology repository and offers a set of web services to enable semantic access to biological data. Querying can be performed over data annotated with the underlying ontologies or containing terms from those ontologies. A query is enriched by additional information using an automated reasoner. Additionally, a service provides semantic search over PubMed.

In the context of the ONKI project the NOR (Normalized Access to Ontology Repositories) [36] approach has been proposed to offer a solution for making ontology repositories universally accessible. The goal is to provide a common search endpoint to different ontology repositories. This approach is not restricted to formal ontologies, but can be applied to all kinds of concept collections with useful identifiers for the Semantic Web (e.g. Wikipedia or Geonames). The NOR approach is based on: (1) a common representation of ontology concepts and (2) a simple API for searching and accessing ontology repositories. The API consists of a concept lookup, a concept search and a metadata endpoints.

OntoCat[13] [2] is a programming interface to query heterogeneous ontology repositories such as Bioportal and OLS, or locally user-specified files. OntoCat implements a wrapper for each repository that enables a uniform access, for example, via a set of REST web services.

Existing terminology systems offer functionalities that are either too general to fit the needs of multiple users (e.g. Finto) or too specific to fit the needs of their own end users (e.g. Aber-OWL).

The systems reviewed shown in Tab. 1 only partly cover the requirements defined in Sect. 2.2. For example, the only solution that offers access to informal terminologies is the NOR system. However, it requires the terminologies invoked to additionally implement and deliver a common representation of their output, which is, in most cases, an implausible solution.

Nearly all the systems reviewed offer hierarchy-oriented services that can be used for query expansion. The Aber-OWL framework is the only solution that offers reasoning services for semantic data access, but does not offer the possibility of combining ontology content with annotations for performing semantic queries. Furthermore, all the systems aim at delivering terms for semantic annotation. The usefulness of these terms for specific applications needs is, however, not guaranteed, as this depends on the choice of the ontology used for annotations and the possibility of using those annotations in combination with the content of ontologies.

None of the existing approaches provide access to the knowledge that can be inferred from the ontologies through reasoning, unless the ontologies have been uploaded in an inferred form, in which case, a limited amount of inferred knowledge will be available.

These insights motivated our decision to set up our own system – the GFBio TS – that is introduced in the next section.

4 A service for a semantic-aware infrastructure

The Terminology Service (TS) extends the GFBio Data infrastructure into a semantic-aware infrastructure. It delivers the backbone terminological services for more efficient data discovery and integration, as well as improved data analysis. In this section, we present the basic design principles of the TS. These basic design principles are based on the requirements in Sect. 2.2.

[11] www.ontobee.org

[12] www.aber-owl.net

[13] www.molgenis.org/wiki/OntocatStart

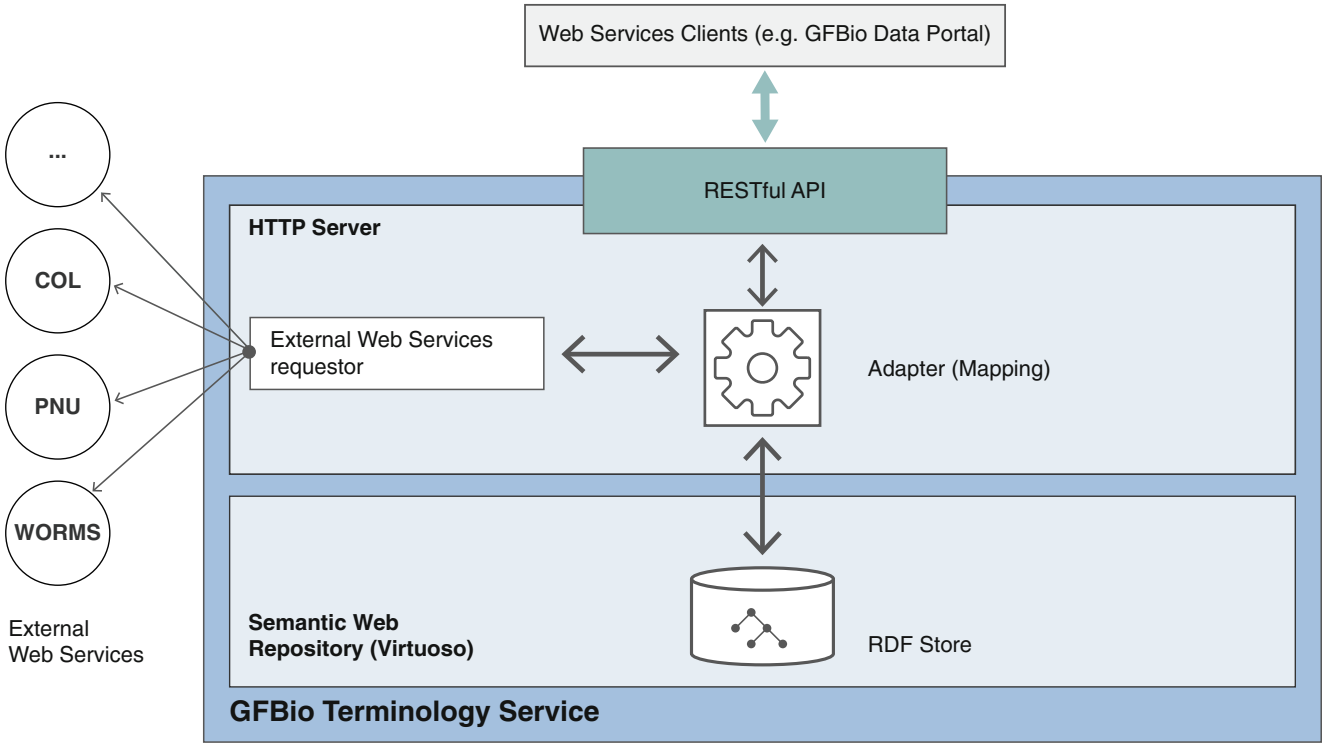

Fig. 2 The GFBio Terminology Service architecture

The general GFBio TS architecture is shown in Fig. 2. It provides a web services interface that serves as a generic access point to heterogeneous terminological resources. In our context the term "terminology" refers to any terminological resource – this can be a formal ontology, a taxonomy, or any useful source of Semantic Web compliant collections of terms (e.g. locations available via Geonames). The TS offers a universal access to the various types of terminologies in a uniform and transparent manner. It delivers unified results enabling computational access to the semantic content of the ontologies in combination with the hierarchical and lexical information held in taxonomies.

These terminologies are either internally hosted or accessed via their remote web services. Internal terminologies are stored in a local Semantic Web repository (Virtuoso[14]). The TS offers additionally access to internally managed terminologies via a Linked Data interface and a SPARQL[15] endpoint. External terminologies must be registered at the TS; they are remotely accessed via a web service requestor. The set of taxonomies registered at the TS include global initiatives like the Catalogue of Life (COL) [28] or the World Register of Marine Species (WoRMS) [1], and taxonomic services from our project partners like the Bacterial Nomenclature Up-to-Date (PNU) [24].

The key component of the TS is the adapter that enables the mapping of both internal Semantic Web terminologies and external terminological resources into a common output format (cf. the gear wheel in Fig. 2).

We describe the content and the implemented API of the TS in the following two sections, then its application in GFBio-related use cases is highlighted.

4.1 Available Terminologies

Unlike existing terminology repositories (cf. Sect. 2.2), the GFBio TS does not aim at providing access to all terminologies of the domain of interest, but only the ones useful in the GFBio context. The task of terminology mobilization is carried out in close connection with our project partners. The types of terminologies (ontologies and taxonomies) we have included so far and also their usefulness in the GFBio context are described in the following.

A large set of biological ontologies has been made available and is used widely in different data-intensive applications. The usefulness of an ontology for specific application needs depends on different criteria. In our context, and in collaboration with our GFBio partners, we evaluated existing ontologies based on their coverage, their added value for data discovery and data presentation, and their quality and development status. The ontologies selected are stored in a local repository for multiple reasons: (1) enable the integration and interlinking of local ontologies; (2) provide

[14] virtuoso.openlinksw.com

[15] www.w3.org/TR/rdf-sparql-query

a home to smaller project-related ontologies and assistance for their creation and maintenance; (3) offer an efficient access via a Semantic Web repository for query expansion by storing the ontologies in a pre-reasoned form; (4) allow for a full access to the ontologies to specific applications where processing is performed using a reasoner.

Storing local copies of the ontologies requires an additional effort to provide updates when new versions are available. A semiautomated verification mechanism based on a web crawler has been set up to check periodically for new versions.

Taxonomies of biological organisms are central resources in biological studies and biodiversity management. A large number of taxonomies exist that are widely used in different projects, but these are often not in a Semantic Web-compliant format. Most of these taxonomies provide stable identifiers that can be used in Semantic Web applications. They are also a great source of hierarchical information that can be exploited in inferences and lexical information, such as common names and synonyms. Some also provide links to similar taxa in different taxonomies making them Linked Data compliant.

Taxonomies are usually accessible via a set of web services and can be easily accessed programmatically. In order to be easilly integrated in specific application scenarios, some taxonomies have been translated into ontologies. The NCBITaxon[16] ontology, for instance, is an automatic translation of the NCBI taxonomy database [12] into an OWL[17] ontology. It is maintained and updated by the OBO (Open Biomedical Ontologies) foundry [30], a consortium of ontology developers concerned with providing a family of interoperable biomedical ontologies. The translation reproduces faithfully all of the content of the source database. It makes it easier to use the taxonomy in Semantic Web applications.

The main drawback we see when considering the translation over querying the taxonomy web services is related to the ontology updates. The NCBI Taxonomy is updated with approximately 500 taxa every week and about 2,000 every month. An OBO administrator manually triggers the releases of the OWL version. According to the homepage[18], the builds are performed every two or three months. This means that up to 8,000 taxa are missing between the ontology releases. Furthermore, the ontological nature of taxa has been discussed in many existing works [29, 32, 34] and limitations of integrating taxonomies into a an ontological framework have been highlighted [13].

The main advantage of using the ontology version would be the ability to perform reasoning, but the NCBITaxon ontology is mainly a description of the taxonomy hierarchy, which does not make it very expressive. Furthermore, as our main goal is to perform ontology-based data access, the hierarchical information delivered by the web services would be sufficient to perform the desired inferences.

From the GFBio users' perspective, this means that newly added taxa will be available, as we will be accessing the latest version of the taxonomy. Additionally, scientists can access the desired data in a more efficient way, as the TS enables query expansion and ontology-based access capabilities for taxonomies as well.

4.2 Normalized API for Accessing Terminologies

Access to the GFBio TS is provided via a RESTful web service. All terms and terminologies can be accessed via a common interface, regardless of whether they are hosted internally or externally. The Terminology Service can be accessed using its public API. The service output is delivered in JSON[19], XML[20], CSV[21] or JSON-LD[22] format. The service endpoints are grouped into four categories: metadata services, search services, information services and hierarchy-oriented services. The API documentation is available at:

http://terminologies.gfbio.org/developer_section/api.html

External terminologies maintain their own schema and deliver their specific attributes, for example the Catalogue of Life (COL) uses "name" for a taxon label while the World Register of Marine Species (WoRMS) uses "CombinedName". For those attributes, we define a Semantic Web compliant common attribute and its corresponding URI:

http://terminologies.gfbio.org/terms/worms-schema/CombinedName

Internal terminologies' attributes correspond to properties defined inside the ontologies. Those are already in a Semantic Web format. In order to achieve a harmonized output, a mapping between Semantic Web attributes and the ones returned by external web services is needed. Thus, we defined a common schema for the TS output. A part of this schema is depicted in the right side of Fig. 3. A mapping to this schema is required for every underlying terminology or connected external service in order to represent the service

[16] The NCBI taxonomy is a curated classification and nomenclature for all of the organisms in the public sequence databases

[17] Web Ontology Language, www.w3.org/OWL

[18] build.berkeleybop.org/job/build-ncbitaxon

[19] JavaScript Object Notation (www.w3.org/TR/html-json-forms/)

[20] Extensible Markup Language (www.w3.org/XML/)

[21] Comma Separated Values (tools.ietf.org/html/rfc4180)

[22] JSON for Linking Data (www.w3.org/TR/json-ld/)

```
@prefix gfbiots: <http://terminologies.gfbio.org/terms/ts-schema/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix omv: <http://omv.ontoware.org/2005/05/ontology#> .
@prefix gfbio: <http://terminologies.gfbio.org/terms/ontology/> .
@prefix col: <http://terminologies.gfbio.org/terms/col-schema/> .

col:name rdfs:subPropertyOf gfbiots:label .
col:rank rdfs:subPropertyOf  gfbiots:rank .
col:taxonStatus rdfs:subPropertyOf gfbiots:status .
col:taxonConceptUuids rdfs:subPropertyOf gfbiots:externalID .
```

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix gfbiots: <http://terminologies.gfbio.org/terms/ts-schema/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<http://terminologies.gfbio.org/terms/ts-schema> a owl:Ontology ;
  dc:title "The GFBio Terminology Server Schema vocabulary" .

gfbiots:label a rdf:Property ;
  rdfs:isDefinedBy <http://terminologies.gfbio.org/terms/ts-schema/> ;
  rdfs:label "label" ;
  rdfs:subPropertyOf rdfs:label ;
  rdfs:range rdfs:Class ;
  rdfs:range rdfs:Literal ;
  rdfs:comment "The label of the term." .

gfbiots:rank a rdf:Property ;
  rdfs:isDefinedBy <http://terminologies.gfbio.org/terms/ts-schema/> ;
  rdfs:label "rank" ;
  rdfs:range rdfs:Class ;
  rdfs:range rdfs:Literal ;
  rdfs:comment "The taxonomic rank of the term." .
```

Fig. 3 Mapping of the COL schema (excerpt) to the TS schema (excerpt)

results in a common format. Attributes that have no correspodance in the TS schema are assigned a URI and stored in their underlying schema. Fig. 3 shows a part of the COL schema (left side) defining its attributes as sub-properties of the corresponding ones in the general TS schema (right side). For instance, the COL attribute "name" is mapped to the TS attribute "label". The schema for each external terminology is defined when the corresponding service is connected, once the mapping defined any changes in the underlying database at the data level is automatically reflected in the API output.

The schemas are rather stable, but dealing with changes at the attributes level is an issue that will occur only in extreme cases. The schemas and the mappings are maintained in the triple store and the properties defined are accessible via a linked data interface and used in the JSON-LD output of the REST API.

The adapter accesses the schema and performs the mapping of all attributes defined into the common TS schema to serialize the output delivered to the web service's clients. An application or component accessing the TS API will get a comprehensible set of attributes and does not need to deal with the specificities of the multiple formats. This makes the TS not only suitable for accessing diverse and heterogeneous terminologies via a single API, but also for combining them.

4.3 Implemented Use Cases

The TS[23] has been running since June 2014. Other components of the GFBio infrastructure build upon its services for enabling semantic capabilities. We report three foundational use cases where the GFBio TS has been applied successfully. We are conducting multiple studies related to the use cases, we describe the ongoing and planned evaluations for each one.

Use Case 1: Semantic recommendation of biodiversity datasets Researchers face a real challenge while using data portals to access relevant data due to the large amount and highly heterogeneous nature of the data sets available. Even formulating the right query is a hard task in that context.

In order to help scientists in that process, a semantically enhanced recommender system for the biodiversity domain has been developed in the context of the GFBio project [25]. The recommender aims to improve data discovery of the highly heterogeneous data sets available based on the semantic tools provided by the TS. In particular, the portal search engine invokes the TS to perform related terms suggestions for expanding user queries. It also provides context and visualizations of the terms considered based on their terminological definition and their relations with other terms. This helps researchers improve and better focus their query, reducing the set of results to the best fitting ones.

In order to evaluate the added value of the semantic recommendation, we conducted an analysis on collected user queries logs. A first step consisted of analyzing the coverage of the actual set of TS terminologies and the possibility to extend it. Next, we plan to compare the results obtained with and without TS support. We are collecting information about user behaviour towards queries results and the collected data will be analysed in terms of precision and recall. We consider a result to be useful if the user accesses the corresponding dataset and downloading the data is an even higher indicator of usefulness.

Use Case 2: Personalization of terminologies access for GFBio data management platforms Data management and archiving processes in GFBio are distributed among partners using a range of different software plat-

[23] terminologies.gfbio.org

forms. Presently, each of these data management platforms handles vocabularies used for data standardization and annotation locally, usually by means of a software module dedicated to the storage, maintenance, and versioning of controlled vocabularies.

The modularized Diversity Workbench (DWB) [33] is a virtual research environment for multiple scientific purposes with regard to the management and analysis of life sciences data. The DWB consumes different vocabularies, such as the PESI taxonomy (Pan-European Species-directories Infrastructure) [21] and the Diversity Taxon Names (DTNtaxonlists[24]), a self-maintained list of domain-specific terms (e.g. checklists, taxon reference lists, red lists).

Accessing the TS API reduces the burden of dealing with local versions of the vocabularies as well as the sources specificities. Moreover, the DWB can use the TS as a layer for specific personalized functionalities, for example, PESI consists of five different embedded databases which cannot be addressed separately in the original service, but this is possible using a specific parameter of TS API.

Considering the virtual environments such as the Diversity workbench, the intoduction of terminological support at the data capture level facilitates the annotation task by providing the researcher with comprehensive terms from the given terminologies. Application scenarios are built in collaboration with developers and users in order to enrich semantically their local functionalities.

Use Case 3: Transfer of a data standards into a semantic-aware format The ABCD (Access to Biological Collection Data) standard [19] has been developed as an XML-schema providing the grammar for a strictly hierarchical organization of collection and observational data. ABCD is widely used as a data transfer format (for example between data providers and data aggregators and portals) but it also serves as a reference model for concepts required in the context of collection data processing in general.

The upcoming version ABCD 3.0[25] will move from the strict and monolithic XML Schema towards an ontological approach, allowing individual concepts to be referenced and reused more easily. It addresses shortcomings of the former versions: (1) ABCD-concepts were identified by their xpath within the ABCD schema. For an improved integration with the emerging semantic-aware infrastructures concepts need to be identified by persistent URIs, and (2) Knowledge about the relations between concepts in ABCD are hidden in the schema hierarchy and not machine readable. For example, an ABCD parent-child structure might represent a sub-class or a substring relation.

[24] www.diversitymobile.net/wiki/DTN_Taxon_Lists_Services

[25] abcd.biowikifarm.net

Using the GFBio TS improves access to ABCD 3.0 by enabling researchers to search for ABCD concepts for annotating biodiversity information at both the data and the metadata level. Furthermore, knowledge engineers can integrate ABCD-concepts in ontologies used in the context of GFBio, and software developers can define application schemes composed of ABCD concepts and used for specific data storage or data transfer scenarios.

5 Current state, limitations and future work

The growing number of data-intensive and interdisciplinary research projects call for improved mechanisms enabling: (1) The discovery of relevant data sets for research purposes, and (2) the integration of data from heterogeneous resources.

Scientists put a lot of effort into developing data standards for sharing information within their respective communities and providing varying levels of semantic interoperability via vocabularies ranging from simple lists of terms, hierarchical mostly XML-based formats to fully featured ontologies. Integration and discovery can be achieved by means of a joint data model into which research data have to be mapped. This is the case in the EUROPEANA approach for cultural and natural history collection objects [20]. We believe, however, that this approach will fail when applied to the highly diverse landscape of biological research and related domain data.

A more promising approach is the harmonization of existing standards by means of a semantic-aware platform integrating existing vocabularies and mediating access to internal and externally maintained standards via an optimized API for annotation and search functions over a distributed data-archiving facility. The GFBio TS has been initiated and intensively developed to fulfill this perspective. An initial set of terminologies is now available via the TS API. These were selected in connection with the developed use cases with a high concern about the quality of their content and the possibility of interlinking and integrating them in an efficient manner. Additional functionalities are added on demand, always in connection with the components actually connected to the TS.

The current use cases highlighted a number of limitations, some of which are specific to our application and others are well-known and have been tackled intensively in the literature. The main limitation resides in the lack of interoperability between the terminologies considered. Interoperability between biological ontologies and taxonomies is discussed extensively [13, 17, 22, 30] and, despite the efforts of the OBO foundry towards developing interoperable ontologies, an efficient strategy is still missing. In order to cope with this limitation, we started developing a higher

level ontology that will serve to clarify the interrelations between different terminologies. The corresponding use case is described later in this section. Another challenge when using existing terminologies is that they are not designed to fit multiple needs and visions. A basic issue we detected is how researchers perceive, for example, definitions or synonyms introduced by others.

A continuous effort is being undertaken to enrich the technical backbone of the GFBio TS, on the one hand, and to improve its content in terms of the terminologies included and connected, on the other hand. The basis for such a process is a constant communication with participating archiving facilities and scientific communities in order to identify their needs and integrate relevant vocabularies and ontologies, and organize the development of additional terminologies.

In particular, two additional use cases are under development in collaboration with our GFBio partners. In the first use case, an high level application ontology is being developed. It will enable interoperability between the various ontologies available by defining higher level links between them. The ontology will serve mainly as a basis for annotations and automated faceted search generation.

Changing the whole search infrastructure to a Semantic Web-oriented solution is not a viable solution in certain application scenarios. Query expansion fails in that case too, considering the size of some of the terminologies. The second use case was introduced to cope with these specific limitations and offers a solution improving search results. We are planning an extension at the data level where data is not only annotated with the corresponding terms from the terminologies but also with additional information derived from the terminologies (e.g. key broader terms).

6 Conclusion

Technology has transformed scientific research practice and has led to new directions in research. At the same time, it is an enabler for greater collaboration between disciplines [26]. We introduced the GFBio TS that extends the GFBio Data Portal by a semantic-aware infrastructure. This extension enables researchers to work together despite their cross-disciplinary research areas. Similar attempts, for example, have been undertaken by the European Commission. The project aims to create a European Open Science Cloud that will allow the storage, sharing and reuse of data across disciplines and borders.

We described existing and planned use cases that build upon the Terminology Service. All realized use cases support researchers at different levels in their research practice: Firstly, when searching for datasets, secondly, when using up-to-date terminologies in their virtual research environments, and finally, when modeling biological data to improve the data exchange. All these use cases and all the future ones are driven by the idea that semantic technologies should provide valuable services to the users on a level where the user does not need to have any knowledge about semantics. The complexity of the service should be hidden by ensuring a high API and UI usability.

References

1. WoRMS Editorial Board (2016) World Register of Marine Specie. http://www.marinespecies.org. Accessed 2016-04-25
2. Adamusiak T, Burdett T, Kurbatova N, Joeri van der Velde K, Abeygunawardena N, Antonakaki D, Kapushesky M, Parkinson H, Swertz MA (2011) Ontocat – Simple Ontology Search and Integration in Java, R and Rest/Javascript. BMC Bioinformatics 12(1):1–12
3. Atkins D, Droegemeier K, Feldman S, Garcia-Molina H, Klein M, Messerschmitt D, Messina P, Ostriker J, Wright M (2003) Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the Blue-Ribbon Advisory Panel on Cyberinfrastructure. National Science Foundation, Washington, DC
4. Authmann C, Beilschmidt C, Drönner J, Mattig M, Seeger B (2015) VAT: A System for Visualizing, Analyzing and Transforming Spatial Data in Science. Datenbank-Spektrum 15(3):175–184
5. Baader F, Lutz C, Suntisrivaraporn B (2006) CEL – A Polynomial-Time Reasoner for Life Science Ontologies. Springer, Berlin, Heidelberg
6. Berendsohn W, Döring M, Geoffroy M, Glück K, Güntsch A, Hahn A, Kusber WH, Li J, Röpert D, Specht F (2003) The berlin model: a concept-based taxonomic information model. In: MoReTax – Handling Factual Information Linked to Taxonomic Concepts in Biology. BfN, Schriftenreihe Vegetationskunde, vol 39
7. Ciardelli P, Kelbert P, Kohlbecker A, Hoffmann N, Güntsch A, Berendsohn WG (2009) The EDIT Cyberplatform for Taxonomy and the Taxonomic Workflow: Selected Components. In 39. Jahrestagung der Gesellschaft für Informatik e.V. (GI). GI, Lübeck, Germany, pp 625–638
8. Deegan (née Clark) JI, Dimmer EC, Mungall CJ (2010) Formalization of Taxon-Based Constraints to Detect Inconsistencies in Annotation and Ontology Development. BMC Bioinformatics 11(1):1–10
9. Côté RG, Jones P, Apweiler R, Hermjakob H (2006) The Ontology Lookup Service, a Lightweight Cross-Platform Tool for Controlled Vocabulary Queries. BMC Bioinformatics 7(1):1–7
10. Diepenbroek M, Glöckner FO, Grobe P, Güntsch A, Huber R, König-Ries B, Kostadinov I, Nieschulze J, Seeger B, Tolksdorf R, Triebel D (2014) Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (gfbio). 44. Jahrestagung der Gesellschaft für Informatik. GI, Stuttgart, Germany
11. Euzenat J, Shvaiko P (2013) Ontology Matching, 2nd edn. Springer-Verlag, Heidelberg (DE)
12. Federhen S (2012) The NCBI Taxonomy Database. Nucleic Acids Res 40(Database issue):D136–D143
13. Franz N (2011) Biological Taxonomy and Ontology Development: Scope and Limitations. Biodivers Informatics. doi:10.17161/bi.v7i1.3927
14. Gerlach R, Blaa D, Chamanara J, Hohmuth M, Navabpour N, Thiel S, König-Ries B (2015) Bexis 2 – A Platform for Managing Heterogeneous Biodiversity Data and Projects. TDWG Annual Conference

15. Hevner AR, March ST, Park J, Ram S (2004) Design Science in Information Systems Research. Mis Q 28(1):75–105
16. Hey T, Tansley S, Tolle KM et al (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery vol. 1. Microsoft research, Redmond, WA
17. Hoehndorf R, Dumontier M, Oellrich A, Rebholz-Schuhmann D, Schofield PN, Gkoutos GV (2011) Interoperability Between Biomedical Ontologies Through Relation Expansion, Upper-Level Ontologies and Automatic Reasoning. PLOS ONE 6(7):1–9
18. Hoehndorf R, Slater L, Schofield PN, Gkoutos GV (2015) Aber-Owl: A Framework for Ontology-Based Data Access in Biology. BMC Bioinformatics 16(1):1–9
19. Holetschek J, Dröge G, Güntsch A, Berendsohn WG (2012) The ABCD of Primary Biodiversity Data Access. Plant Biosyst 146(4):771–779
20. Isaac A, Haslhofer B (2013) Europeana Linked Open Data – data.europeana.eu. Semant Web 4(3):291–297
21. de Jong Y, Kouwenberg J, Boumans L et al (2015) Pesi – A Taxonomic Backbone for Europe. Biodivers Data J 3:e5848
22. Köhler S, Bauer S, Mungall CJ, Carletti G, Smith CL, Schofield P, Gkoutos GV, Robinson PN (2011) Improving Ontologies by Automatic Reasoning and Evaluation of Logical Definitions. BMC Bioinformatics 12(1):1–8
23. Kuśnierczyk W (2008) Taxonomy-Based Partitioning of the Gene Ontology. J Biomed Inform 41(2):282–292
24. Leibniz Institute DSMZ (2016) Prokaryotic Nomenclature Up-To-Date. http://www.dsmz.de/bacterial-diversity/prokaryotic-nomenclature-up-to-date
25. Löffler F, Sateli B, Witte R, König-Ries B (2014) Towards semantic recommendation of biodiversity datasets based on linked open data. In: Proceedings of the 26th GI-Workshop Grundlagen von Datenbanken, vol. 1313. Bozen-Bolzano, Italy, pp 65–70
26. Meyer ET, Schroeder R (2015) Knowledge Machines: Digital Transformations of the Sciences and Humanities. MIT Press, Cambridge, MA
27. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MD, Chute CG, Musen MA (2009) Bioportal: Ontologies and Integrated Data Resources at the Click of a Mouse. Nucleic Acids Res 37(Web-Server-Issue):170–173
28. Roskov Y, Abucay L, Orrell T, Nicolson D, Flann C, Bailly N, Kirk P, Bourgoin T, DeWalt R, Decock W, De~Wever A (eds) (2016) Species 2000 & ITIS Catalogue of Life, 25th March 2016. www.catalogueoflife.org/col (Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-8858)
29. Schulz S, Stenzhorn H, Boeker M (2008) The ontology of biological taxa. In: Proceedings 16th International Conference on Intelligent Systems for Molecular Biology (ISMB) Toronto, Canada, July 19–23, 2008. pp 313–321
30. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S (2007) The Obo Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nat Biotechnol 25(11):1251–1255
31. Suominen O, Pessala S, Tuominen J, Lappalainen M, Nykyri S, Ylikotila H, Frosterus M, Hyvönen E (2014) Deploying national ontology services: From onki to finto. In: Proceedings of the Industry Track at the International Semantic Web Conference 2014. CEUR Workshop Proceedings
32. Thau D, Ludäscher B (2007) Reasoning About Taxonomies in First-Order Logic. Ecol Inform 2(3):195–209 (Meta-information systems and ontologies. A Special Feature from the 5th International Conference on Ecological Informatics ISEI5, Santa Barbara, CA, Dec. 4–7, 2006 Novel Concepts of Ecological Data Management S.I.)
33. Triebel D, Hagedorn G, Jablonski S, Rambold G (eds) (1999) Diversity Workbench – A virtual research environment for building and accessing biodiversity and environmental data. http://www.diversityworkbench.net
34. Tuominen J, Laurenne N, Hyvönen E (2011) Biological names and taxonomies on the semantic web – managing the change in scientific conception. In: The Semanic Web: Research and Applications – 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, Proceedings, Part II. pp 255–269
35. Viljanen K, Tuominen J, Hyvönen E (2009) Ontology libraries for production use: The finnish ontology library service onki. In: Proceedings of the 6th European Semantic Web Conference
36. Viljanen K, Tuominen J, Mäkelä E, Hyvönen E (2012) Normalized access to ontology repositories. In: Proceedings of the Sixth International Conference on Semantic Computing (IEEE ICSC 2012). IEEE Press, Washington, DC
37. Xiang Z, Mungall C, Ruttenberg A, He Y (2011) Ontobee: A linked data server and browser for ontology terms. In: Proceedings of the 2nd International Conference on Biomedical Ontology Buffalo, NY, USA, July 26–30, 2011