FREIE UNIVERSITÄT
BERLIN

The GFBio Terminology Service: enabling a research data management beyond data heterogeneity

Naouel Karam, Robert Harald Lorenz, Claudia Müller-Birn

TR-B-17-01
March, 2017

FACHBEREICH MATHEMATIK UND INFORMATIK
SERIE B • INFORMATIK

Abstract

A primary goal of a research infrastructure for data management should be to enable efficient data discovery and integration of heterogeneous data. The German Federation for Biological Data (GFBio) was envisioned by this goal. The basic component, that enables such interoperability and serves as a backbone for such a platform, is the GFBio Terminology Service (GFBio TS). It acts as a semantic platform for accessing, developing and reasoning over terminological resources within the biological and environmental domain. A RESTful API gives access to these terminological resources in a uniform way regardless of their degree of complexity and whether they are internally stored or externally accessed through their web services. Additionally, a set of widgets with an intrinsic API connection are made available for an easy integration in applications and web interfaces. Based on the requirements of the GFBio partners, we describe the added value that is provided by the GFBio Terminology Service with practical scenarios but also, what challenges we still face. We conclude by describing our current activities and future developments.
Keywords: Research data infrastructure, Interoperability, Terminology repository, Semantic Web, RESTful API, Widgets

1 Introduction

Research practice has become more data-intensive over the last few decades, and this development is visible across many research disciplines. However, the sharing of research data beyond disciplinary borders is still a challenge. Thus, a research infrastructure for data management should allow for an efficient data integration and therefore, the discovery of heterogeneous research data.

The German Federation for Biological Data (GFBio) pursues this goal. GFBio aims at providing a data management platform and data archiving solutions for data capture, annotation, indexing, searching and storage in the area of biological and environmental research. The GFBio Data Portal[1] integrates existing data infrastructures such as PANGAEA[2] into the GFBio Repository Network.

Data generated in biodiversity and ecology research are extremely heterogeneous and pertaining to different scientific disciplines using various methods and technologies. The situation is further complicated by different understandings of employed terms within different scientific domains. Developing interoperability and harmonizing data by using standards and terminological resources are crucial for data mobilization, integration, and discovery in the GFBio context.

The core component that enables this interoperability and serves as a backbone for the GFBio infrastructure is called the GFBio Terminology Service[3] (GFBio TS) [15]. The GFBio Terminology Service acts as a semantic platform for accessing, developing, and reasoning over terminological resources. The GFBio TS focuses on integrating and giving access to terminologies developed by

[1] www.gfbio.org

[2] www.pangaea.de

[3] terminologies.gfbio.org

project partners as well as external terminologies defined and maintained by related communities. These terminologies can range from simple term lists to complex ontologies. Based on the requirements of the GFBio community, the Terminology Service provides access to over 20 terminologies so far, where GFBio partners have contributed 10 terminologies. A well-defined RESTful API gives access to all terminologies in a uniform way regardless of their degree of complexity and whether they are internally stored or externally accessed through their web services.The services provided by the GFBio TS can also be integrated easily within existing web applications with the help of widgets, which are small applications with limited functionality. We developed two exemplary widget prototypes so far: a term visualization and a search widget.

We will explain the advantage of using semantic technologies for data management and highlight the utility of the Terminology Service by practical use cases of semantically enhanced components. More specifically, we will differentiate between four main usage scenarios developed so far: *Explore*, *Access*, *Download* and *Contribute*. In the *Explore* scenario, researchers can reuse ontologies that are interesting for their research. In the *Access* scenario developers can use information in ontologies programmatically to provide semantically enriched applications and web services. In the *Download* scenario, information from the ontologies can be retrieved and stored to a local information system. In the *Contribute* scenario, we consider that scientists can store their terminologies in the TS to access all provided services automatically. Finally, we discuss existing challenges in this field that are often in the social-technical context.

2 A common infrastructure for biological data

GFBio [12] is developing an infrastructure to enable biological and environmental scientists to share and discover their data more efficiently. It 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 [21], the Bexis system [13] or the EDIT Platform for Cybertaxonomy [10]. Infrastructure is being extended by a semantic component that ensures, 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 research infrastructure of GFBio, consisting of four main components.

The GFBio Data Portal integrates existing data infrastructures into the GFBio Repository Network (bottom in Fig. 1). The latter comprises amongst others molecular data (EMBL-EBI[4]), environmental data (PANGAEA[5]), as well

[4] `www.ebi.ac.uk`

[5] `www.pangaea.de`

as natural history and culture collection data (e.g. MfN[6], DSMZ[7] and SNSB[8]).

The data provided by portal users are indexed and semantically enriched, which provides a meaning for the data. Understanding the data allows scientists to analyze and visualize them, for example by using the GFBio VAT System (Visualization, Analysis & Transformation system) [9]. The possibility to enrich data with semantic information is provided by a fourth component - the GFBio Terminology Service. The semantic meaning is enabled by the provision and interlinking of ontologies and taxonomies.

There are existing systems providing a comparable terminology service. These systems can be either full platforms for terminology management [19, 11, 20, 14, 24] or frameworks for accessing terminologies [8, 23]. We defined a set of requirements related to our project needs and analyzed to what extent existing systems meet those requirements [15]. One requirement was to be able to integrate well established taxonomies like the World Register of Marine Species (WORMS)[9] or the Catalogue of Life (COL)[10]. Those taxonomies are widely used in the domain for annotating species for example and they are a source of valuable hierarchical information. None of the existing systems integrates such type of terminologies. Additional requirements relate to our project philosophy where we aim to provide tools and inference mechanisms specifically tailored to our GFBio partners. The derived insights motivated our decision to set up our own system – the GFBio Terminology Service – that is introduced in the next section.

3 The Terminology Service

We describe in this section the main building blocks of the GFBio Terminology Service. First, we introduce the basic concepts and define the term *terminology* in our context. Then, we present the general architecture of the GFBio TS.

3.1 Basic Concepts

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 a geographical database like Geonames[11]). It encompasses several meanings ranging from simple lists of terms to semantically rich ontologies. Unfortunately, there are currently no commonly accepted definitions of the different terminology types (in the biological domain) which leaves room for variation causing them to be used interchangeably depending on the context.

[6] www.naturkundemuseum.berlin

[7] www.dsmz.de

[8] www.snsb.mwn.de

[9] www.marinespecies.org

[10] www.catalogueoflife.org

[11] www.geonames.org

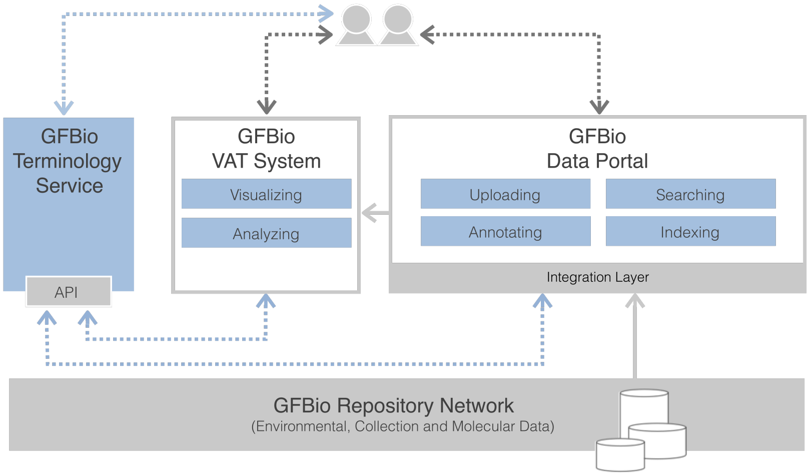


Figure 1: The GFBio components.

We introduce our agreed terminology formality levels, with differing levels of specifications going from the most informal to the most formal level as described in Figure 2. The different levels are illustrated by the term *water*[12] that is extracted from the CHEBI ontology[13] and depicted in Figure 3.

GFBio distinguishes between five different types or formality levels of terminologies. The less formal level contains of a *Controlled Vocabulary*. It is the simplest type of terminology that consists of a finite list of terms consisting only of labels without definitions or hierarchical ordering. Based on the example, only the label *water* is part of the terminology.

The next formality level is *Glossary*. It is a list of term labels that includes an informal definition of their meaning in natural (human-readable) language additionally. Since information expressed in natural language is typically not unambiguous, these specifications are not yet adequate for further processing by computer agents. In a glossary, the definition of the term *water* is partnered by its label.

In a *Taxonomy*, a term is a compound of a label, a definition and hierarchical information, e.g., by *is-a* relationships, thus providing additional semantics in the relations between the terms which can be interpreted by computer agents. The hierarchical structure depicted in Figure 3 would be part of a taxonomy describing the term *water*.

A *Thesaurus* is a controlled vocabulary connected via relations between the terms expressing hierarchies (e.g., *narrower*/*broader term*), associations (e.g., *related term*), or synonym relationships. In the example, a thesaurus contains

[12] `http://purl.obolibrary.org/obo/CHEBI_15377`
[13] `www.ebi.ac.uk/chebi`

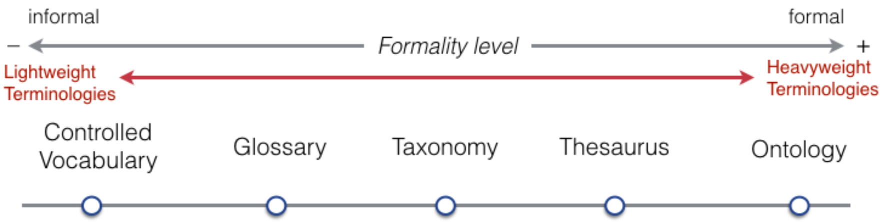
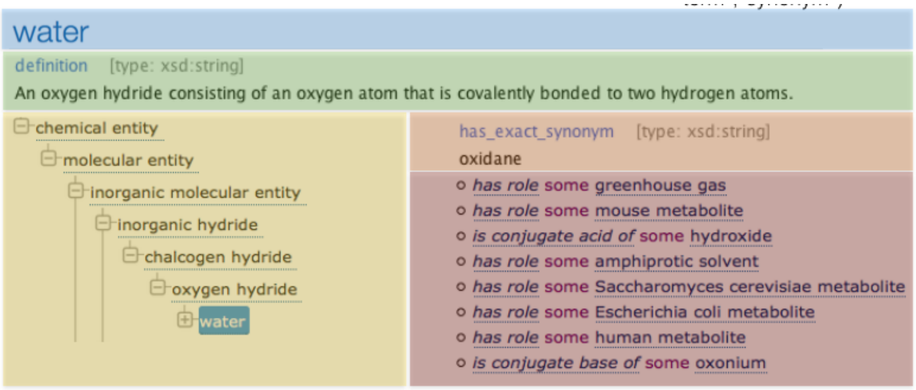


Figure 2: GFBio agreed terminology formality levels.

the information about the synonym *oxidane* of the term *water*.

The most formal terminology is an *Ontology*. A term consists of all the information provided at the lower levels augmented with complex relationships, allowing an unambiguous interpretation of terms and relationships according to logic-based rules. In our example, an ontology would contain the whole spectrum of relations we already considered in the other levels and additional complex or user defined relations like *has_ role* and *is_ conjugat_ base_ of*.



water

definition [type: xsd:string]

An oxygen hydride consisting of an oxygen atom that is covalently bonded to two hydrogen atoms.

- chemical entity
  - molecular entity
    - inorganic molecular entity
      - inorganic hydride
        - chalcogen hydride
          - oxygen hydride
            - water

has_exact_synonym [type: xsd:string]

oxidane

- *has role* some greenhouse gas
- *has role* some mouse metabolite
- *is conjugate acid of* some hydroxide
- *has role* some amphiprotic solvent
- *has role* some Saccharomyces cerevisiae metabolite
- *has role* some Escherichia coli metabolite
- *has role* some human metabolite
- *is conjugate base of* some oxonium

Figure 3: Excerpt of the definition of the term water of the CHEBI ontology.

3.2 The Terminology Service Architecture

The general architecture of the Terminology Service is shown in Figure 4. In March 2017, the Terminology Service gives access to over 20 terminologies that have been requested by the GFBio partners so far. Those terminologies are either internally stored in a Semantic Web repository or remotely accessed via their web services. Internal terminologies are stored in a local RDF[14] store in a Semantic Web compliant format such as OWL[15] or SKOS[16]. Internal termi-

[14] www.w3.org/RDF
[15] www.w3.org/OWL
[16] www.w3.org/2004/02/skos/

nologies can be accessed directly via a Linked Data interface and a SPARQL[17] endpoint. The included terminologies are well established ones like the CHEBI ontology for example or ontologies provided by the GFBio community like the KINGDOM[18] ontology, describing a GFBio agreed list of species kingdoms. Table 1 lists the actual status of included terminologies with information about their type, storage and if they are provided by GFBio partners.

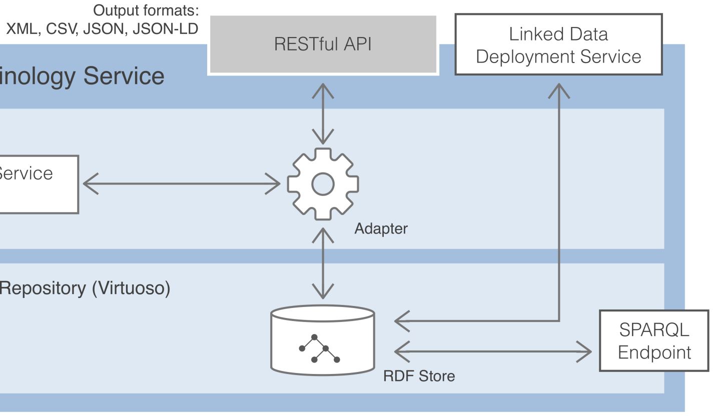

Output formats:
XML, CSV, JSON, JSON-LD
RESTful API
Linked Data Deployment Service
GFBio Terminology Service
HTTP Server
External Web Service Requestor
Adapter
Semantic Web Repository (Virtuoso)
RDF Store
SPARQL Endpoint

Figure 4: The GFBio Terminology Service architecture.

The Terminology Service software is being developed using Java based on the Jena[19] Semantic Web framework. We implemented an external web service requestor for obtaining seven external taxonomies (such as the Catalogue of Life). A key component of the TS is the adapter component (cf. the gear wheel in Figure 4) that enables the schema mapping of both internal and external terminological resources into a common output format. We defined a common schema for the Terminology Service output. A mapping to this schema is required for every underlying terminology or connected external service in order to achieve a harmonized API output. For instance, the COL attribute *name* is mapped to the GFBio TS attribute *label*. Thus, all terms and terminologies can be accessed via a common interface (the RESTful API), regardless of whether they are hosted internally or externally. The service output is delivered in four formats: JSON, XML, CSV, and JSON-LD. This interface allows developers who are not familiar with semantic technologies or Linked Data to easily access the provided terminologies efficiently.

[17] terminologies.gfbio.org/sparql
[18] terminologies.gfbio.org/terms/KINGDOM
[19] jena.apache.org

Table 1: List of terminologies included in GFBio Terminology Service (ON = Ontology, TAX = Taxonomy).

Type	Storage	Language	Acronym	Name	GFBIo
ON	internal	OWL	BCO	Biological Collections Ontology	no
ON	internal	OWL	PATO	Phenotypic Quality Ontology	no
ON	internal	OWL	CHEBI	Chemical Entities of Biological Interest Ontology	no
ON	internal	OWL	RECORDBASIS	GFBio Agreed Vocabulary for RecordBasis	yes
ON	internal	OWL	ENVO	Environment Ontology	no
ON	internal	OWL	OBOE	Extensible Observation Ontology	no
ON	internal	OWL	KINGDOM	GFBio Agreed Vocabulary for Kingdoms	yes
ON	internal	OWL	QUDT	Quantity, Unit, Dimension and Type	no
ON	internal	OWL	SWEET	Semantic Web for Earth and Environment Technology Ontology	no
ON	internal	SKOS	ISOCOUNTRIES	ISO 3166 Countries and Subdivisions	yes
ON	internal	SKOS	LIT_I	The lithologs rock names ontology for igneous rocks	yes
TAX	internal	OWL	NCBITAXON	National Center for Biotechnology Information (NCBI) Organismal Classification	no
ON	internal	OWL	BOHLMANN	Bohlmann Ontology	yes
TAX	internal	OWL	ORIBATIDA	Oribatida Ontology	yes
TAX	internal	OWL	THYSANOPTERA	Thysanoptera Ontology	yes
TAX	internal	OWL	TRICHOPTERA	Trichoptera Ontology	yes
TAX	external	?	DTNtaxonlists_SNSB	Regionalised and Domain-specific Taxon Lists	yes
TAX	external	?	COL	Catalogue Of Life	no
TAX	external	?	PNU	Prokaryotic Nomenclature up-to-date	yes
TAX	external	?	ITIS	Integrated Taxonomic Information System	no
ON	external	?	GEONAMES	The GeoNames geographical database	no
TAX	external	?	PESI	Pan-European Species directories Infrastructure	no
TAX	external	?	WORMS	World Register of Marine Species	no

4 Accessing the Terminology Service

The GFBio Terminology Service can be accessed either through its common interface - the RESTful API[20] – or using widgets we provide; these are small web applications with limited functionality which allow for user interactions. We describe in the following both ways to access the GFBio TS.

4.1 The Terminology Service API

The RESTful API of the Terminology Service can be used programmatically by connecting the service to web services such as the GFBio Data Portal or the VAT (cf. Figure 1) or other applications. At the moment the API provides 14 endpoints that are organised into terminology-specific, term-specific, search, and hierarchy-oriented endpoints. Details about the calls signatures, the parameters as well as example calls can be found in the API documentation section on our website (`terminologies.gfbio.org`). In the following, we describe each category briefly and provide a tabular description for each endpoint.

4.1.1 Terminology-specific endpoints

The four terminology-specific endpoints, which are described in Table 2, provide information on terminologies like the list of available terminologies and their metadata, such as the name, description and creation date.

Table 2: Terminology-specific endpoints

Endpoint	Description
List all terminologies	Returns the list of all available terminologies of the GFBio TS. The result set contains the name, acronym (terminology-id), short description and URI of each terminology.
Get the information about a terminology	Returns the information about a terminology given its acronym. The result set contains the URI, acronym, name, description, domain, ontology language, creation date and expressivity of the terminology.
Display the metrics of a terminology	Returns the metrics of a terminology. They are of two types, statistical metrics like the number of classes or properties and quality-control metrics like the number of classes without a label or a definition.
Display the metadata of a terminology	Returns the metadata contained in the RDF file of a given terminology.

[20] Application Programming Interface

4.1.2 Term-specific endpoints

Term-specific endpoints relate to particular terms from the terminologies. One can list all terms of a specific terminology, query the information about a term or get the list of its synonyms (c.f. Table 3).

Table 3: Term-specific endpoints

Endpoint	Description
Get all terms of a terminology	Returns the list of terms of a given terminology. The result set contains the label and URI of each term.
Get information about a term	Returns the information about a term, like its label and definition, given its URI.
Get the synonyms of a term	Returns the synonyms of a term given its URI.

4.1.3 Search endpoints

Two search endpoints are provided, the first one returns all terms corresponding to a query string, the second is implemented for suggesting terms while users are typing.

Table 4: Search endpoints

Endpoint	Description
Search	The search looks inside labels, synonyms, common names, acronyms and abbreviations. Possible search types are exact, included or regular expression based term matches, the default search returns terms that correspond exactly to the searched string. Further parameters can be used to restrict the search, a detailed list can be found in our API documentation.. The result set contains the label, URI, description, rank, kingdom, source terminology and synonyms or common names of each matching term.
Suggest	Returns all terms containing a given string, limited to 15 suggestions by default. This endpoint can be used for suggesting terms in a dropdown menu for example.

4.1.4 Hierarchy-oriented endpoints

Hierarchy-oriented endpoints return information relative to the position of a term in the hierarchical structure of the terminology. Broaders and narrowers terms of a given term can be returned as well as the complete hierarchical path up to the top of the hierarchy.

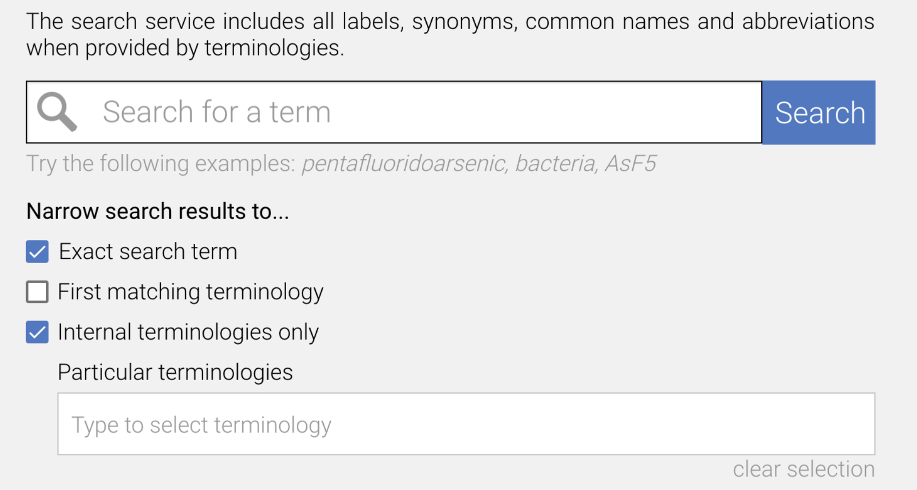
Table 5: Hierarchy-oriented endpoints

Endpoint	Description
Get narrower terms	Retrieves the term(s) that are one level narrower than a given one. The result set contains the URI and label of the narrower term(s).
Get all narrower terms	Retrieves all terms that are narrower of a given one including each possible path to the leaves of the hierarchy. The result set contains the URI and label of all narrower terms.
Get broader terms	Retrieves the term(s) that are one level broader than a given one. The result set contains the URI and label of the broader term(s).
Get all broader terms	Retrieves all terms that are broader of a given one including each possible path to the top. The result set contains the URI and label of all broader term(s).
Get the top hierarchy	Retrieves the hierarchical path to the top level for a given term. The result set contains the URI, label and the direct broader terms URIs of all terms in the hierarchy.

4.2 The Terminology Service Widgets

The Terminology Service provides widgets – that are components, "chunks of web page" or small applications – intended to be used within web pages. The widgets deliver a restricted functionality, often for just one purpose, like displaying data or providing an interface. Typically, a widget contains a mixture of HTML, CSS and JavaScript where the complexity is ideally hidden to make it as easy as possible for developers to integrate the widgets to their application or website with little configuration and programming skills needed. All of our widgets use the Terminology Service API and thus, users can quickly expand their local service with all the functionalities provided by the GFBio TS API. Our goal is to provide reusable and easy to use widgets to be integrated and reused easily with none or little knowledge in web development. Furthermore, the widgets are licensed under an open source licence and will be published openly on Github soon. At the moment, we prototypically implemented two widgets: a term visualisation and a search widget. In the following, we take the latter as an example, to show the methodological approach for developing widgets.

The search widget allows users to search for terms from terminologies to determine their usefulness for their work, e.g. for annotating research data in the GFBio Data Portal. Before developing this widget, we examined 13 services which provide search functionalities in the same or related fields as ours. The majority (6) of the examined services allowing to look for classes (terms) in particular ontologies or vocabularies (Cropontology[21] [2], Finto[22] [20], Ontobee[23] [24], Aber-Owl[24] [14], Bioportal[25] [19], OLS[26] [11]). The latter three are capable of searching for ontologies as well. Three services (Biosharing[27] [16], VEST[28] [7], ANDS[29] [1]) looking for vocabularies, ontologies, policies or standards only and four (Datacite[30] [3], Dryad[31] [4], F1000research[32] [5], Vertnet[33] [6]) are for searching scientific papers and data resources. The presentation of the search interface differs a lot. From very simple interfaces to advanced ones with many search options and filter functionalities. We examined design criteria like the overall size of the widget, the position and layout of the submit button, the placeholder text of the search bar, the availability and presentation of advanced search functionalities and help sections. Our research suggests the following main considerations which resulted in the prototypical design depicted in Figure 5.


The search service includes all labels, synonyms, common names and abbreviations when provided by terminologies.

Try the following examples: *pentafluoridoarsenic, bacteria, AsF5*

Narrow search results to...

☑ Exact search term

□ First matching terminology

☑ Internal terminologies only

Particular terminologies

Figure 5: Screenshot of the GFBio TS search widget prototype.

The development process included the investigation of a widget scaffold where the objective was twofold: the development process for further widgets

[21] www.cropontology.org

[22] www.finto.fi

[23] www.ontobee.org

[24] www.aber-owl.net

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

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

[27] http://biosharing.org

[28] http://vest.agrisemantics.org/vocabularies

[29] http://vocabs.ands.org.au

[30] www.datacite.org

[31] www.datadryad.org

[32] http://f1000research.com

[33] http://portal.vertnet.org

Table 6: Criteria for widget development

Search input
- Placeholder text mentions *what* can be found instead of e.g. *where* the search is performed - Searching in progress is displayed in input field to show waiting time in user's gaze
Submit button
- Button has the same height as the search input - Strong background color shows possible action
Advanced search options
- Interaction elements for all available options the GFBio TS API provides (*internal_ only*, *match_ type*, *first_ hit*, particular terminologies) - Options are always shown and not hidden [17] - Default advanced search options narrowing down search results instead of broadening [22] - Text has same font type and size as default text [22] - Possibility to apply the search refinements after the search was done as well (filter-like functionality)
Help
- Additionally explanations are not needed for easy-to-use and self-explaining interface - Runnable examples help the users
General
- Responsive design helps developers to control the size of the embedded widget to their special needs - Wording signalizes actions to be done (*Narrow down*, *Try example*, *Search for*) - Non-technical wording is important because users will most likely be non-computer scientists

should be simplified and standardized as well as the the process for developers to integrate our widgets into their websites. We then investigated three services (Google[34], Twitter[35] and ANDS[36] [1]) providing customized widgets.

[34] `https://developers.google.com`

[35] `https://dev.twitter.com`

[36] `http://vocabs.ands.org.au`

With some kind of guidance users are able to click through options on the website to receive customized HTML code and references to JavaScript and style files to be embedded on their own website. As customisation is planned but not implemented yet, our goal is to deliver one JavaScript and one CSS file to be integrated in the users HTML via the corresponding HTML markups. Because our widgets will deliver a broad spectrum of functionality the scaffold consists next to the way how developers integrating it, of the module design pattern, used libraries, a shared layout file and partly shared functions.

5 Using the Terminology Service within GFBio

Currently, the GFBio community uses the Terminology Service within four main scenarios. Each scenario has been defined and developed in cooperation with the GFBio partners. Each partner provides discipline and context specific requirements on the GFBio TS. The development of these use cases is an ongoing process and further use cases will be provided in the near future.

In the *Browse* scenario users, i.e. researchers, can peruse terminologies that are interesting for their research. For this, the visualization widget provides term details and shows a term's position within a tree structure, if the terminology is a taxonomy or in a graph structure, if the terminology is an ontology. In the GFBio Data Portal the visualization can be used in the research data submission process. When annotating the data in the submission process, the user can can easily browse term details and explore existing term relations by type to identify those terms that describe their data best.

In the *Access* scenario developers can use information in terminologies programmatically to provide semantically enriched web services based on the GFBio TS. In the GFBio Data Portal, the TS allowed for developing a semantic search service for research data. Based on query expansion, the original search term is extended by related terms from different terminologies in order to provide a more comprehensive overview on existing research data.

In the *Consume* scenario, information from terminologies of the GFBio TS can be retrieved and stored to a local information system. In the GFBio context, this is needed for data management within small and medium scale projects that are carried out by virtual research environments such as BExIS [13] and Diversity Workbench [21]. In these contexts, the provided metadata from the terminologies of the TS can be pre-processed to support the data annotation process locally.

In the *Contribute* scenario we consider that researchers or data curators can store their individual terminologies in the GFBio TS. Instead of developing their own terminology management system, this will allow them to access all services provided by the TS easily. For example, in the GFBio context, the partners have contributed ten terminologies so far. Those terminologies are either internally stored like the KINGDOM ontology or connected as external web services like the DTN Taxon Lists Services or the Prokaryotic Nomenclature Up-to-Date and interna. In GFBio, the mobilization of community-relevant terminologies is supported by an internal process. The terminology owner can register the terminology in the internal wiki and in collaboration with the terminology curator the needed metadata are provided. If the metadata are complete, a terminology is manually integrated into the GFBio TS.

6 Current activities and next steps

We introduced the GFBio TS that extends the GFBio infrastructure with semantic capabilities. This extension enables researchers to share their data despite their heterogeneous nature. After presenting the project context and the basic concepts, we described the Terminology Service general architecture and the way to accessing and integrating it using its public interface or via a set of downloadable widgets.

We described concrete use cases that support researchers at different levels in their research practice, for example, when searching for datasets or when using up-to-date terminologies in their virtual research environments.

At the moment, a high level application ontology, the GFBio ontology is being developed. It will enable interoperability between the various terminologies available by defining higher level links between them. Moreover, this ontology will serve mainly as a basis for annotations and automated faceted search.

We are working on the integration of the semantic annotation tool neonion [18] within the GFBio context. The aim is to allow scientists to annotate information in scientific texts with terminologies coming from the GFBio TS, and thus, research results and research data can be more closely connected.

The interoperability issue is due to different understandings of terms within different scientific domains or to the use of different labels to refer to the same term. This issue can be solved by annotating data with terms from the Terminology Service. Data can still be annotated using equivalent terms coming from different terminologies. In order to ensure interoperability the underlying terminologies should be interlinked. We are developing a semi-automated mapping service and interface based on a combination of matching algorithms.

The GFBio TS is continuously updated to meet partners needs. A set of tools is being developed to support terminologies selection based on query and text analysis as well as tools for transforming terminologies from text and tabular forms into a Semantic Web compliant format.

References

[1] Australian national data service website. `www.ands.org.au`. Accessed: 2017-01-06.

[2] Crop ontology curation and annotation tool – 2011 generation challenge programme, bioversity international as project implementing agency. `www.cropontology.org`. Accessed: 2017-01-06.

[3] Datacite website. www.datacite.org. Accessed: 2017-01-06.

[4] Dryad digital repository website. www.datadryad.org. Accessed: 2017-01-06.

[5] F1000Research website. http://f1000research.com. Accessed: 2017-01-06.

[6] VERTNET: Distributed databases with backbone website. http://portal.vertnet.org. Accessed: 2017-01-06.

[7] VEST / AgroPortal - Map of standards website. http://vest.agrisemantics.org/vocabularies. Accessed: 2017-01-06.

[8] Tomasz Adamusiak, Tony Burdett, Natalja Kurbatova, K. Joeri van der Velde, Niran Abeygunawardena, Despoina Antonakaki, Misha Kapushesky, Helen Parkinson, and Morris A. Swertz. Ontocat – simple ontology search and integration in java, r and rest/javascript. *BMC Bioinformatics*, 12(1):1–12, 2011.

[9] Christian Authmann, Christian Beilschmidt, Johannes Drönner, Michael Mattig, and Bernhard Seeger. VAT: A system for visualizing, analyzing and transforming spatial data in science. *Datenbank-Spektrum*, 15(3):175–184, 2015.

[10] Pepé Ciardelli, Patricia Kelbert, Andreas Kohlbecker, Niels Hoffmann, Anton Güntsch, and Walter G. Berendsohn. The EDIT cyberplatform for taxonomy and the taxonomic workflow: Selected components. In *In 39. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Lübeck, Germany*, pages 625–638, 2009.

[11] Richard G. Côté, Philip Jones, Rolf Apweiler, and Henning Hermjakob. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7(1):1–7, 2006.

[12] Michael Diepenbroek, Frank Oliver Glöckner, Peter Grobe, Anton Güntsch, Robert Huber, Birgitta König-Ries, Ivaylo Kostadinov, Jens Nieschulze, Bernhard Seeger, Robert Tolksdorf, and Dagmar 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*.

[13] Roman Gerlach, David Blaa, Javad Chamanara, Martin Hohmuth, Nafiseh Navabpour, Sven Thiel, and Birgitta König-Ries. Bexis 2 – a platform for managing heterogeneous biodiversity data and projects. In *TDWG Annual Conference*, 2015.

[14] Robert Hoehndorf, Luke Slater, Paul N. Schofield, and Georgios V. Gkoutos. Aber-owl: a framework for ontology-based data access in biology. *BMC Bioinformatics*, 16(1):1–9, 2015.

[15] Naouel Karam, Claudia Müller-Birn, Maren Gleisberg, David Fichtmüller, Robert Tolksdorf, and Anton Güntsch. A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data. *Datenbank-Spektrum*, 16(3):195–205, 2016.

[16] Peter McQuilton, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, Milo Thurston, Allyson Lister, Eamonn Maguire, and Susanna-Assunta Sansone. Biosharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database*, 2016:baw075, 2016.

[17] Peter Morville and Jeffery Callender. *Search Patterns - Design for Discovery.* O'Reilly, 2010.

[18] Claudia Müller-Birn, Tina Klüwer, André Breitenfeld, Alexa Schlegel, and Lukas Benedix. neonion: Combining human and machine intelligence. In *18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14-18, 2015, Companion Volume*, pages 223–226, 2015.

[19] Natalya Fridman Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne D. Storey, Christopher G. Chute, and Mark A. Musen. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web-Server-Issue):170–173, 2009.

[20] Osma Suominen, Sini Pessala, Jouni Tuominen, Mikko Lappalainen, Susanna Nykyri, Henri Ylikotila, Matias Frosterus, and Eero Hyvönen. 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.

[21] Dagmar Triebel, Gregor Hagedorn, Stefan Jablonski, and Gerhard Rambold (eds.). Diversity Workbench - A virtual research environment for building and accessing biodiversity and environmental data, 1999 onwards.

[22] Steve Turbek. Advancing advanced search. `http://boxesandarrows.com/advancing-advanced-search`, 2008. Accessed: 2017-01-13.

[23] Kim Viljanen, Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen. Normalized access to ontology repositories. In *Proceedings of the Sixth International Conference on Semantic Computing (IEEE ICSC 2012)*. IEEE Press, September 2012.

[24] Zuoshuang Xiang, Chris Mungall, Alan Ruttenberg, and Yongqun He. 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*, 2011.