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.

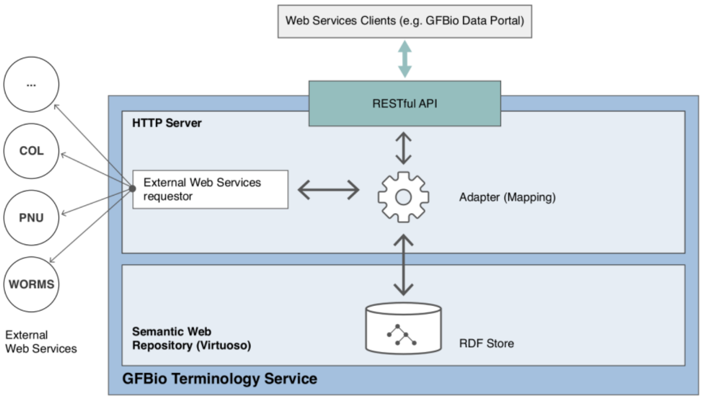
Web Services Clients (e.g. GFBio Data Portal)
RESTful API
HTTP Server
External Web Services requestor
Adapter (Mapping)
...
COL
PNU
WORMS
External Web Services
Semantic Web Repository (Virtuoso)
RDF Store
GFBio Terminology Service

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)

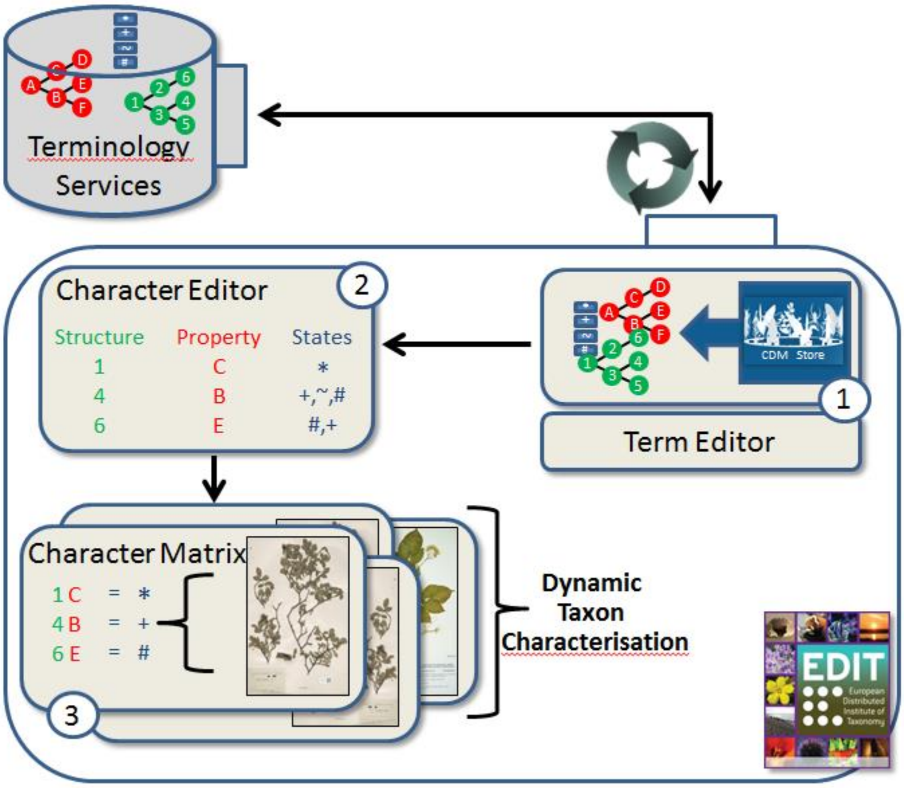

Fig 2 The EDIT platform uses the API of the terminology service to integrate the terminology services into three applications: 1) the term editor which allows editing on a synced copy of the ontology, 2) the character editor where the user defines taxon specific term hierarchies for structures, properties 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.