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

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

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

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

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

1 Introduction

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

[3] www.dfg.de
[4] www.gfbio.org

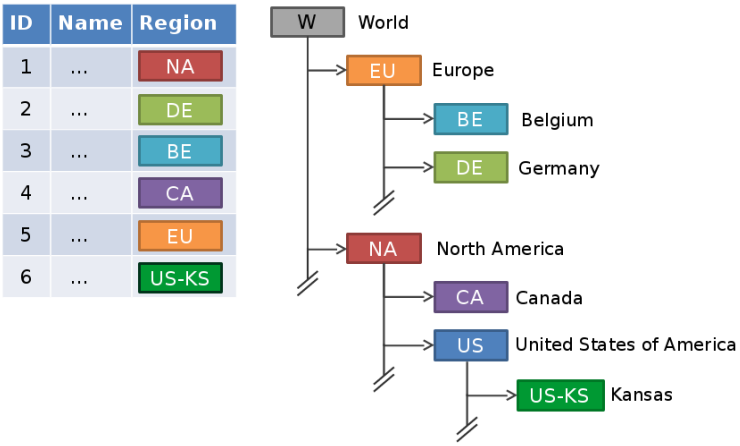

Fig. 1: A simple example of a geographic classification as it is used in research data and as it exists on its own with its definitions and connections.

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

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

[5] www.re3data.org
[6] www.datacite.org

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

2 Context and Related Work

2.1 What are Terminologies

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

2.2 What is GFBio and the GFBio Terminology Service

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

---

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

[8] terminologies.gfbio.org

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

[10] www.catalogueoflife.org

[11] www.marinespecies.org

[12] www.geonames.org

2.3 Related Initiatives

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

3 Terminology Preparation Steps

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

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

[14] www.openrefine.org

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

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

4 Advantages of accessible Research Terminologies

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

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

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

5 Conclusion

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

References

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