Extending a SKOS-based taxonomy catalog with collaborative features and an interface to provide terminologies to describe research data with interdisciplinary, semantic concepts

André Langer, Bach Tran and Martin Gaedke

Professorship for Distributed and Self-organizing Systems,
Chemnitz University of Technology, Germany

Publishing research data in the World Wide Web is typically done by uploading scientific files into a research data repository. Additional meta information can be provided, which is then used to improve the discoverability of this research dataset. However, search operations and filters are mainly keyword-based and commonly result in additional irrelevant or even missing search results, especially in an interdisciplinary research data sharing context. A semantic, concept-based approach can address this issue by relying on well-established taxonomies and linking similar concepts together. Taxonomy services already exist in different knowledge domains and provide concepts with identifiers in a controlled, quite static and isolated way. Essential features, such as collaboration, linking and integration, are often limited or missing, which are success factors for Web 2.0 applications and services. We therefore envision an interdisciplinary taxonomy service both accessible for humans and applications that can provide research concepts from different domains together with unambiguous identifiers and a flexible API to retrieve and manage available terms.

1 Introduction

In the context of OpenScience, researchers are encouraged to publish their research datasets in common data repositories so that others can find and reuse it. Following the FAIR Principles for scientific data management and stewardship [1], „(Meta)data (shall be provided with) assigned globally unique and persistent identifiers“ (F1). Establishment already took place for persistent identifiers that are relevant for mainly administrative and citation metadata, encompassing approaches such as DOI for unambiguously identifying publications, ORCID for identifying authors, ROR for organizations, and further more [2]. Beside that, standardized vocabularies are increasingly used to provide metadata for

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13733

research publications in a structured way, such as based on OpenAIRE Guidelines for Data Archives [8] (DataCite, DCAT-AP, Dublin Core).

Nevertheless, finding existing relevant research datasets in practice for reuse, replay or repurpose is still a challenge, as search queries have to be formulated carefully [4] and search results have to be reviewed individually, if they are actually related to the current research focus and required research data characteristics, or not, as shown in figure 1.

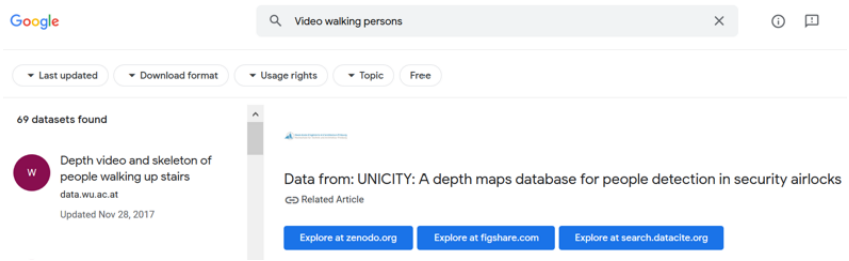

Figure 1: Example research data search query using Google Dataset search.

Reasons for that are, but are not limited to:

1. Filter operations are restricted to general characteristics, such as the download format, usage rights or topic area, as structured metadata descriptions commonly focus on general aspects as the least common denominator among knowledge areas
2. Search operations are commonly based on a keyword-based search within the available metadata title, list of keywords, and natural-language/floating text-based abstract
3. Providing highly structured metadata descriptions with mappings to similar concepts from other disciplines is demanding for users as not many appropriate research data publishing tools exist that consider structured concepts for descriptive purposes.

Especially the last aspect is astounding, as the FAIR principles suggest the user to provide rich metadata (F1) in standardized vocabularies (I2) with accurate and relevant attributes (R1). By using established terminologies with unique concept identifiers, researchers can make sure, that the provided meta information is structured, unambiguous and appropriate for retrieval activities in the future. Several web-based taxonomy catalogs and services already exist providing services for accessing a collection of existing terminologies and concepts. But provided terms might be incomplete, outdated or even contain wrong information. Instead, it would be a benefit if such a knowledge base of research-related concepts can be extended by a broader community, including the creation of missing terms, the categorization of existing concept groups and the interlinking between related concepts.

As part of the PIROL PhD project on Publishing Interdisciplinary Research Over Linked data [4], we currently implement and assess the extension of traditional terminology services with collaborative features. The results will directly contribute to the activities

in the ongoing collaborative research center SFB1410 Hybrid Societies. In this highly interdisciplinary research projects, more than 100 scientists from all main research areas closely work together to investigate the future interaction between people and smart devices, which will also require new methods to publish and discover relevant research data in a structured way.

The rest of the paper is structured in the following way: In section 2, we first elaborate the problem scenario and formulate objectives for improving the interdisciplinary research publishing and discovery process. Existing services are characterized in section Related Work. After that, we illustrate a conceptual architecture for an interdisciplinary, collaborative taxonomy service in section BIRD Concept. Section Conclusion will summarize our idea and describe our next steps.

2 Problem Analysis

When publishing research data, a metadata description should be provided that contains relevant indexable content for any potential stakeholder that might search for this dataset. The scientist carrying out the data description and publishing activity therefore has to carefully describe relevant characteristics from the user's point of view by providing appropriate words as keywords or textual content. In the following, we will focus on an uni-language scenario.

The main challenge from our perspective is, that even a very careful description can be insufficient for the discovery of this research data, because users in the future might use

1. **Generic or specialized terms**: broader class types or specialized words for a certain characteristic
2. **Homonyms**: equal words in a totally different context
3. **Synonyms**: different words than the publisher to express the same characteristic
4. **Weasel Words**: equal words that have a somehow different meaning in their discipline
5. **Unaware terms**: aspects that are applicable to the current data publication but not considered in the description by the publishing user

Aspect 1 can be addressed by relying on taxonomic (hierarchical) relationships and inference possibilities. Aspect 2 is related to the ambiguity of certain word labels and requires a classification or typification of terms with the same name but a different meaning. Aspect 3 and 4 can be taken into consideration by thesauri, that link similar words together. Aspect 5 cannot easily be addressed and is out-of-scope of our investigation.

Semantic technologies already exist that provide basic solutions for these aspects by introducing a uniform representation of knowledge-domain specific terminologies together with persistent unambiguous identifiers that represent a particular concept of the real world ("from strings to things") [6]. The feasibility and strength of such an approach was already shown in industrial application scenarios, such as manufacturing process chains, e-commerce or general-purpose search engines. In a scientific context, taxonomies for a certain knowledge area also already exist, but commonly in a decentralized, independent

or even unstructured availability. Thus, it is not trivial to access existing established identifiers for research-specific concepts, expose a list of known species for a particular concept type or map terms with a similar meaning together [7].

Our proposition therefore emphasizes an approach on how to make scientific concepts in existing research terminologies accessible in an interdisciplinary context by focusing on the following objectives:

OBJ1 Provision of an interdisciplinary semantic taxonomy platform

OBJ2 Ability to import existing taxonomies

OBJ3 Ability to filter for particular concepts, types and identifiers

OBJ4 Ability to collaboratively retrieve, add or update concepts

OBJ5 Ability to tag concepts or group of concepts

OBJ6 Ability to link concepts among different terminologies

The described semantic taxonomy service shall be a collaborative point for research concept information access and interdisciplinary reuse, and primarily designed as a knowledge base for application-based access by research data publishing tools. It can assist a user to describe research data submissions in a highly structured, unobtrusive, transparent way.

3 Related Work

The curation and standardization of relevant terms for a particular knowledge discipline is increasingly demanded for the digitization and exchange in scientific communication. One prominent representative is the initiative National Research Data Infrastructure (NFDI)[1] in Germany, which is currently being established through multiple consortia to offer platforms, services and counseling among all major knowledge disciplines in order to also create a link to international efforts, such as the European Open Science Cloud (EOSC) [8].

Dedicated services to access controlled terminologies are already offered [9] in various representations: *Terminology catalogs* (such as NCBO BioPortal, AgroPortal, gfbio [10]) and ontology collections (such as LOV, AberOWL, ORR, OLS, Ontobee) offer directories to search for existing controlled vocabularies, concepts and relationships within a particular research area. *Authority services* (such as LCSH, MeSH, FAST, EuroVoc) provide general classification concepts. And *knowledge bases* (such as DBpedia, Wikidata and Yago) offer structured data about encyclopedic information for general concepts from the real world [11].

The technical basis to realize these platforms are either dedicated implementations or Open source software projects (such as Skosmos, iQvoc). Especially the last group already makes use of a semantic (RDF) data model with standardized vocabularies (SKOS), however, access is commonly limited to read-only operations. Collaborative approaches

[1] `https://www.nfdi.de/`

are rare and realized in Wiki-based environments (such as based on Semantic MediaWiki), but naturally focus on general-purpose entity collections without a scientific research concept focus and limited internal data organization, import and homogeneous mapping possibilities.

4 Concept

To facilitate the interdisciplinary description of research data on a meta level, essential building blocks are currently missing that can provide structured information about sets of scientific concepts, such as investigated objects, applied methods, used devices, described characteristics or research objectives, and links to similar concepts among research disciplines.

In order to combine the strengths of a dedicated taxonomy registry with collaborative features, previously only known from general-purpose Wikis, we envision the **extension of a semantic base platform** that can set concepts from existing scientific terminologies into a relationship. These terminologies will remain decentralized and independent in their actual location and namespace, but the interdisciplinary taxonomy service will enhance access and mapping, and can collect improvement suggestions. A conceptual architecture is shown in figure 2.

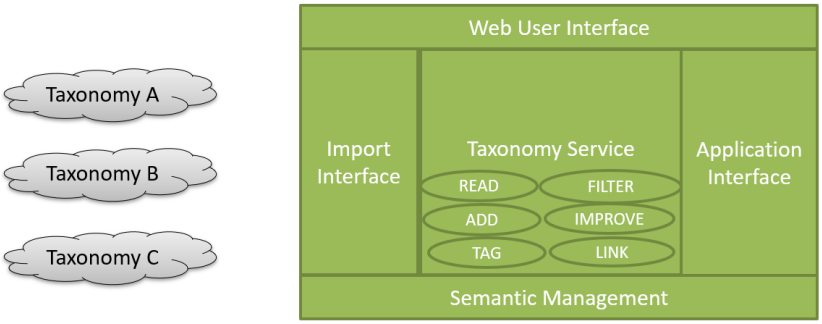

Figure 2: Interdisciplinary research taxonomy service with collaborative extensions.

An **Import Interface** offers a frontend possibility to import an already existing scientific taxonomy into the data corpus of the terminology service. In existing registries, this was either not possible at all, or only manually for administrators in the backend. This is especially relevant for the curation of proposed controlled vocabularies among different disciplines, e.g., across NFDI consortia. The taxonomy has to be represented in a common structured serialization format and is converted into a SKOS representation. Newly added taxonomies can of course go through a review process before being publicly incorporated.

A **Web User Interface** and **Application Interface** is offered to retrieve structured information (also including unambiguous persistent identifiers) about a particular taxonomy or concept. Extended filter possibilities will also allow querying for lists of concepts of a certain type or for synonyms among different terminologies.

An **Add and Improve functionality** allows suggestions for additions and corrections of the existing taxonomical content, as this might not be complete due to the open nature. These operations do not have to be carried out manually via the frontend interface, but can be announced in the background through any data annotation or publishing tool. The suggestions will be set into relationship to a specific taxonomy, but not incorporated in the original, controlled namespace.

A **Tagging component** realizes grouping operations of similar concepts among taxonomies from different disciplines to improve filtering operations for specific types.

Furthermore, a **Linking extension** sets the basis for further link discovery algorithmic operations and inference possibilities along similar research concepts.

5 Conclusion

In this paper, we described our vision for an interdisciplinary taxonomy service providing research concepts in a semantic way, both accessible for humans and machines. It adresses the challenge, that existing general encyclopedic services contain research-related concepts only in an incomplete way with limited categorization and filter possibilities. Controlled taxonomies that standardize scientific concepts and characteristics already exist, but in a decentralized, independent and inhomogeneous way. We have the hypothesis, that bridging research disciplines for research data publishing and discovery is a crowd-based effort to facilitate interdisciplinary research data publishing. Therefore, dedicated taxonomy platforms with additional collaborative features to improve, tag and link similar groups of research concepts have to be realized in order to make them accessible both for humans and research data management applications.

We currently work on establishing such an interdisciplinary taxonomy service originating in the Collaborative Research Center SFB1410 Hybrid Societies. It is built upon Skosmos and allows the collaborative collection, enhancement and integration of terminologies related to human-computer-interaction, overspanning all major research areas. Our next step is to populate it with existing taoxnomies from this knowledge domain and demonstrate its user input assistance in the submission form of research data publishing tools.

Acknowledgements

This work is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) Project ID 416228727 SFB 1410.

ORCID IDs

- André Langer https://orcid.org/0000-0001-7073-5377
- Martin Gaedke https://orcid.org/0000-0002-6729-2912

Bibliography

[1] Mark D. Wilkinson, Michel Dumontier, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. https://doi.org/10.1038/sdata.2016.18

[2] Angela Dappert, Adam Farquhar, Rachael Kotarski, and Kirstie Hewlett. Connecting the persistent identifier ecosystem: Building the technical and human infrastructure for open research. *Data Science Journal*, 16(0):1–16, jun 2017. https://doi.org/10.5334/dsj-2017-028

[3] Pedro Príncipe, Najla Rettberg, Eloy Rodrigues, Mikael K. Elbæk, Jochen Schirrwagen, Nikos Houssos, Lars Holm Nielsen, and Brigitte Jörg. Openaire guidelines: Supporting interoperability for literature repositories, data archives and cris. *Procedia Computer Science*, 33:92–94, 2014. 12th International Conference on Current Research Information Systems, CRIS 2014. URL: https://www.sciencedirect.com/science/article/pii/S1877050914008059, https://doi.org/https://doi.org/10.1016/j.procs.2014.06.015

[4] Hamlet Batista. 7 Reasons Why Search Engines Don't Return Relevant Results 100% of the Time. 2007. https://moz.com/blog/7-reasons-why-search-engines-dont-return-relevant-results-100-of-the-time, Accessed: 2021-05-01.

[5] André Langer. PIROL : Cross-domain Research Data Publishing with Linked Data technologies. In Marcello La Rosa, Pierluigi Plebani, and Manfred Reichert, editors, *Proceedings of the Doctoral Consortium Papers Presented at the 31st CAiSE 2019*, pages 43–51, Rome, 2019. CEUR.

[6] Christopher Erdmann, Natasha Simons, et al. Top 10 FAIR Data & Software Things, 2019. https://librarycarpentry.org/Top-10-FAIR/2019/09/05/linked-open-data/, Accessed: 2021-05-01. https://doi.org/10.5281/zenodo.2555498

[7] Ceri Binding and Douglas Tudhope. Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries*, 17(1):5–21, 2016. https://doi.org/10.1007/s00799-015-0166-y

[8] Javad Chamanara, Angelina Kraft, Sören Auer, and Oliver Koepler. Towards Semantic Integration of Federated Research Data. *Datenbank-Spektrum*, 19(2):87–94, jul 2019. https://doi.org/10.1007/s13222-019-00315-w

[9] Andreas Ledl. Demonstration of the BAsel Register of Thesauri, Ontologies & Classifications (BARTOC). In *NKOS Workshop 2015*, 2015.

[10] 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, nov 2016. URL: `www.naturkundemuseum.berlin`, `https://doi.org/10.1007/s13222-016-0231-8`

[11] André Langer, Christoph Göpfert, and Martin Gaedke. Querying the Semantic Web for Concept Identifiers to Annotate Research Datasets. *Fourteenth International Conference on Advances in Semantic Processing SEMAPRO 2020*, (c):49–55, 2020.