First User Research Workshop on Terminology Services in NFDI4Culture

Lozana Rossenova, Lucia Sohmen, and Kolja Bailly
Open Science Lab, TIB – Leibniz Information Centre for Science and Technology

Introduction

On July 12, 2022, team members from the Open Science Lab, TIB, co-convened a special meeting of the NFDI4Culture LOD Working Group for a workshop focusing on terminology services. Terminologies[1] play an important role across several key focus points of the Task Areas in NFDI4Culture, including TA1's Data enrichment, TA2's Data Standards, and TA5's focus on Knowledge Graphs.

Currently, terminology service tools in use within other NFDI consortia, such as the Ontology Lookup Service used by NFDI4Ing and NFDI4Chem, struggle to accommodate vocabularies and ontologies used in the arts and humanities disciplines due to their typically large size, variety of serialisation formats and complexity of hierarchical relations within the category trees. In the context of TA1 and TA5, members of the Open Science Lab at TIB are prototyping a new service which aims to address the gaps in current service provisions. The motivation to host a dedicated workshop was to ensure that the new service meets the most pressing needs of the NFDI4Culture community in this area.

The workshop featured hands-on exercises with existing terminology services in order to observe how culture community participants interact and make use of such

---

[1] 'Terminologies' is used here as an umbrella term to encompass domain-specific vocabularies, thesauri and formal ontologies, the latter preferably represented through Semantic Web specifications (OWL, SKOS, etc). The FAIR principles explicitly state that a formal and common terminology must be used for knowledge representation. An important task for every NFDI consortium is therefore to identify and align their relevant terminologies within the designated communities and beyond to achieve the broadest possible applicability and acceptance. Because of this there is a dedicated Terminology Service Working Group within the cross-NFDI Section Metadata, and new proposals for TS software solutions are being put forth as part of the ongoing work to establish Base4NFDI.

services. Furthermore, the workshop aimed to identify the most widely used vocabulary and ontology resources in the community and discuss what common issues arise when researchers are browsing or looking up terms in these resources. In the end, the Open Science Lab team presented current service architecture plans and an initial prototype built around a set of speculative user stories. The goal of the workshop was to test the feasibility of these user stories against the real-world scenarios the workshop participants face in their daily research activities.

Twelve diverse culture community representatives took part in the workshop, including researchers, archivists and computer scientists working in the fields of musicology, performing arts, fine art, architecture and design, among others.

Findings from the workshop sessions

In three consecutive tasks, participants were asked to 1) Describe their current workflows; 2) Test some existing services; and 3) Describe their ideal user journeys.

Workshop tasks presented in Miro by Lozana Rossenova (OSL, TIB Hannover). CC-BY.

1) Current workflows and common challenges

The first task asked participants to describe their current workflows step-by-step – what tasks involving terminologies they have to perform on a daily basis. Participants were encouraged to share challenges and success stories, too. They were also asked to share a list of all current terminologies they consult or tools they use to access and work with terminologies.

The resulting text descriptions and subsequent discussions included the following **common scenarios**:

- Researchers and other professional staff have to manually search for a single term across multiple services until they find a match;
- Although there are some known aggregator services, most participants still manually visit individual resource websites to search for terms. Some participants also work with data dumps from individual terminologies and develop their own query scripts customised to the individual resource.
- There is also a lot of research effort required in disambiguation especially with people's names when there is little known other descriptive metadata.
- Very often there are no matching terms in the general resources for a specific field of study, so subject specialists develop their own vocabularies and ontologies;
- In their day to day work advising others on metadata issues, participants also noticed that across the broader culture community there is still a lack of understanding of the value of mapping custom terms to existing terminologies via SKOS (or OWL) mapping properties.
- The mapping work itself is challenging as it requires both nuanced understanding of Semantic Web specifications and SKOS (or OWL) properties, as well as community capacity to garner consensus among niche specialisms.
- There is a lack of streamlined mechanism to contribute back to standard terminologies, if new mappings are created or new terms need to be suggested.
- Once matching terms are discovered, or new mappings are created, the work of inputting term URIs into collection management software still involves many manual steps (even when some workflows are assisted by reconciliation tools like OpenRefine).
- A further common challenge is the need to work with translated terms: many terminologies lack adequate translations to German (and other languages).

Based on these use scenarios the following **key community needs** were identified:

- Avoiding multiple endpoints to search across available standard terminologies;
- Reducing volume of disambiguation-, matching- and mapping-related tasks currently performed manually;
- Publishing custom terminologies following FAIR principles and encouraging the mapping of various degrees of alignment (exactMatch, narrowMatch, broadMatch, etc.) to existing authority control terms via SKOS;
- Allowing for multilingual work (at least full support for German).

The culture community already uses a **range of tools** to assist them in their workflows, and these already work towards addressing some of the issues above. For example:

- xTree, Cocoda, or VocBench for mapping/editing;
- xTree, Skosmos, or Pudel (service-in-the-making) for publishing academic data models and vocabularies;
- OpenRefine for reconciliation (i.e. semi-automated matching of terms against a particular authority control resource), but the available reconciliation service endpoints are limited;
- Bartoc as aggregated search platform for discovery of ontologies and other classification schemas (but not for search of individual terms);
- Wikidata as an (unofficial) aggregation and/or mapping service.

The majority of workshop participants use these **core terminology sources** (in alphabetical order below):

- GeoNames
- Getty Vocabularies (AAT, ULAN, TGN)
- GND (via lobid.org and GND Explorer)
- Iconclass
- VIAF
- Wikidata

Individual users also mentioned using the following terminology sources and registries: CVMA XMP-Metadatenspezification, DiGA Thesaurus, GBIF, IMSLP, LCCN, OBG, RDA Vocabularies, Wortnetz Kultur.

In discussion with participants, the following **conclusions** were drawn during the workshop:

- Although various aggregators exist, a service that provides a single entry point to searching at least the core sources outlined above would be very valuable in reducing manual efforts.
- While services for publishing new terminologies, mapping across new and existing terminology sources, as well as performing reconciliation or entity recognition operations on structured and unstructured data do exist, none of these connect readily to collection management systems, leaving plenty of tasks to be done completely manually.
- Translation requires tackling on a community and maintainer level (i.e. communities of researchers establishing communication channels with the responsible terminology maintainers). This is planned as part of Task Area 2,

Measure 3: 'Cross-linking of specialist vocabularies' in NFDI4Culture which involves activities to address better translation to German of crucial resources such as the Getty's AAT. Still, a service that provides language-specific discovery and matching features would be very important in the context of NFDI4Culture.

2) Testing existing services

During the second task, users were asked to work with three existing terminology services from the cultural field: Finnish Ontology Library Service ONKI, Termennetwerk (Dutch Digital Heritage Network), and GND Explorer. They had to find a term of their choice in at least two authority control sources discoverable via these services, and then to note down positive and negative aspects of the user journey.

Positive observations highlighted:

- The simplicity of the interface design of the Termennetwerk service – resulting in ease-of-use also for beginners: for example, how quickly one could find URIs there;
- Being able to search multiple ontologies at once (as in Termennetwerk);
- The hierarchical display of results in the ONKI service;
- The GND Explorer's performance speed, native German language features and the network view of results (noted for its aesthetic appeal).

Negative observations highlighted:

- Issues with translation (of both terminologies and interface features) because the functionality of the Termennetwerk and ONKI services in English was fairly superficial.
- Lack of domain specific vocabularies in some cases.
- Issues with degree of confidence whether the search results were the right matches for the searched terms.
- Lack of options to filter or facet results in more granular ways (i.e. not just by source of the terminology).
- Difficulty in finding the 'right' entry point (e.g. 'gold spoon' might be found in some terminologies, but others might have it under 'goldsmithing' or just 'household').
- Issues with separating bibliographical and terminological entries in the GND Explorer (e.g. having to click on "Sachbegriff", or sometimes not being able to find a generic term – only its mention in bibliographical entries)
- Issues with visualisation of the results (e.g. when displaying the network view in the GND Explorer, one cannot access the relations (edges), so it is hard to

get an understanding of the overall data model and make further judgements based on that).

Key take-aways from this exercise for a new terminology service developed with the needs of the culture community in mind:

-> Provide full multilingual support.

-> Provide access to domain-specific vocabularies for the culture community.

-> Provide a variety of ways to visualise and display search results.

-> Provide assistance to users in making matching or disambiguation decisions (by a mix of machine-learning techniques, results visualisation approaches, and selection of what information to reveal as part of the search results, e.g. term description, term category tree, term URI, related terms, etc).

3) Ideal user journeys

The last task for which participants were asked to work individually involved diagramming an ideal user journey using some basic graphic elements in Miro (or writing it out in plain text).

The journeys drawn by the participants featured some predictable flows (like searching and finding matching terms), but also some 'wishlist' functionalities, such as the possibility to directly connect the terminology service to one's own collection and/or research database interface and to establish that connection only for relevant vocabularies from the specific domain context. Such a connection would further allow to preview search results including not only term labels, but also definitions, related metadata, and even use-case examples. Crucially, the connection would allow to directly include data from the authority control services (e.g. Iconclass, Getty AAT, GND, VIAF, or even Wikidata) into one's own database (i.e. do 'data extension'[2] via fully or semi-automated workflows). The journeys involving collection management tasks ended with published research datasets which would ideally include data pulled from and/or linked to the authority control services, and would allow further reuse, too.

---

[2] Data extension is the term used by tools such as OpenRefine to refer to data extended from one service e.g. Getty Vocabularies into one's own dataset based on matched terms.
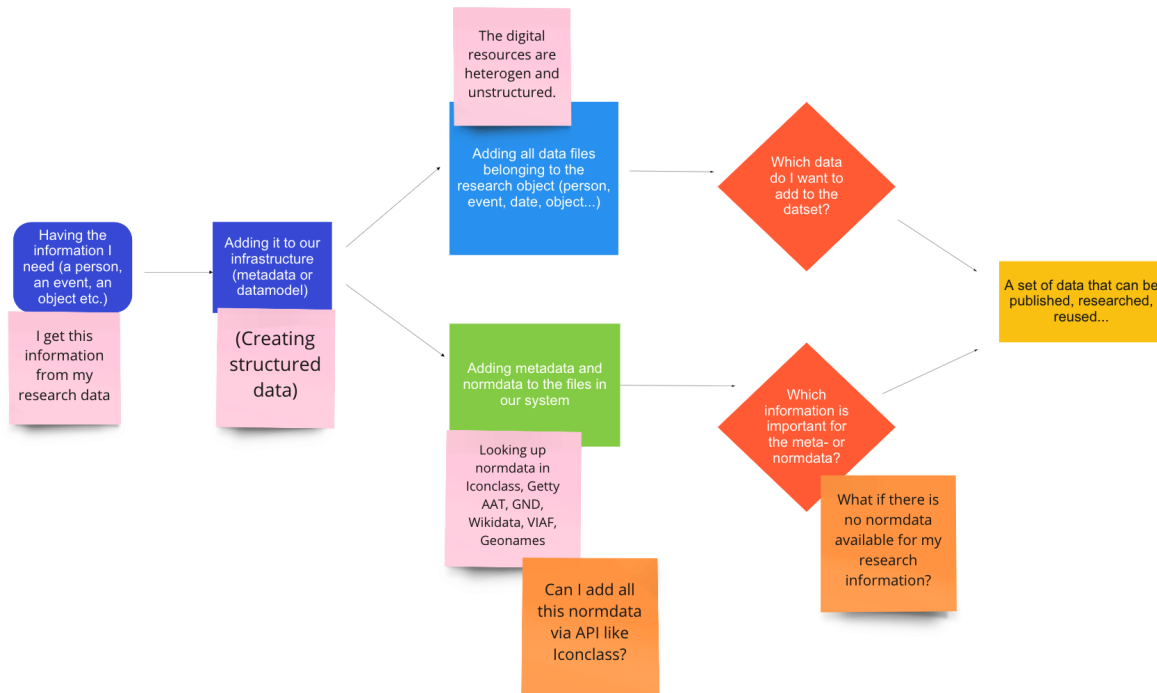
User journey diagramme by Anja Gerber (Berlin-Brandenburgische Akademie der Wissenschaften Corpus Vitrearum Medii Aevi). CC-BY.

Some journeys included data modeling steps and the development of custom terminologies that then were also connected to external authority control sources. Those connections would ideally be bi-directional - i.e. not only connecting bespoke terminologies to standard ones, but also suggesting new terms for inclusion.
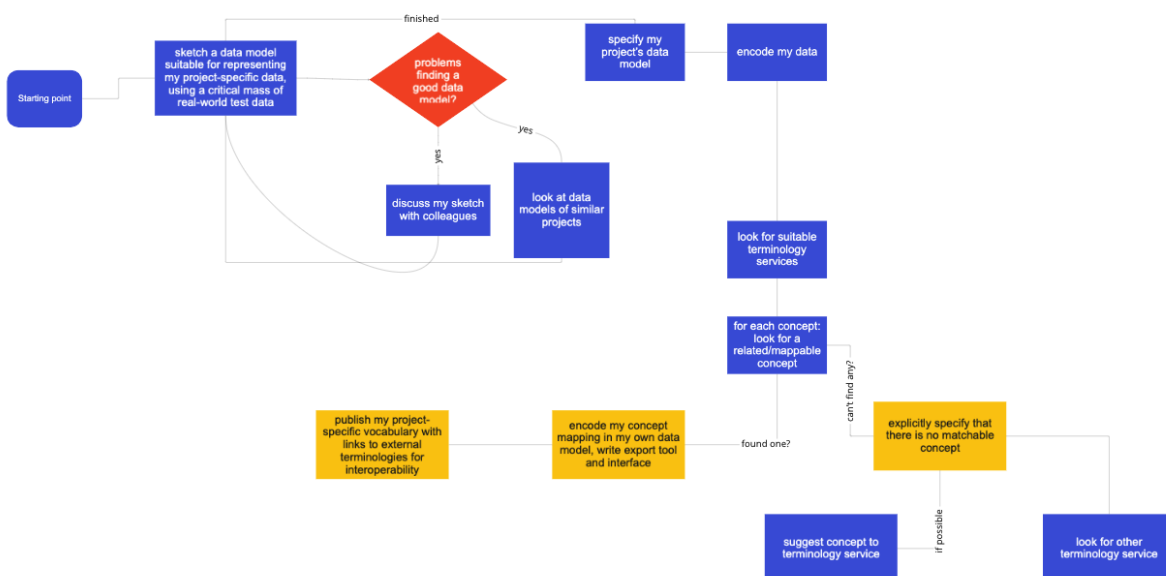
User journey diagramme by Dominik Leipold (Ludwig-Maximilians-Universität München, Institut für Musikwissenschaft). CC-BY.
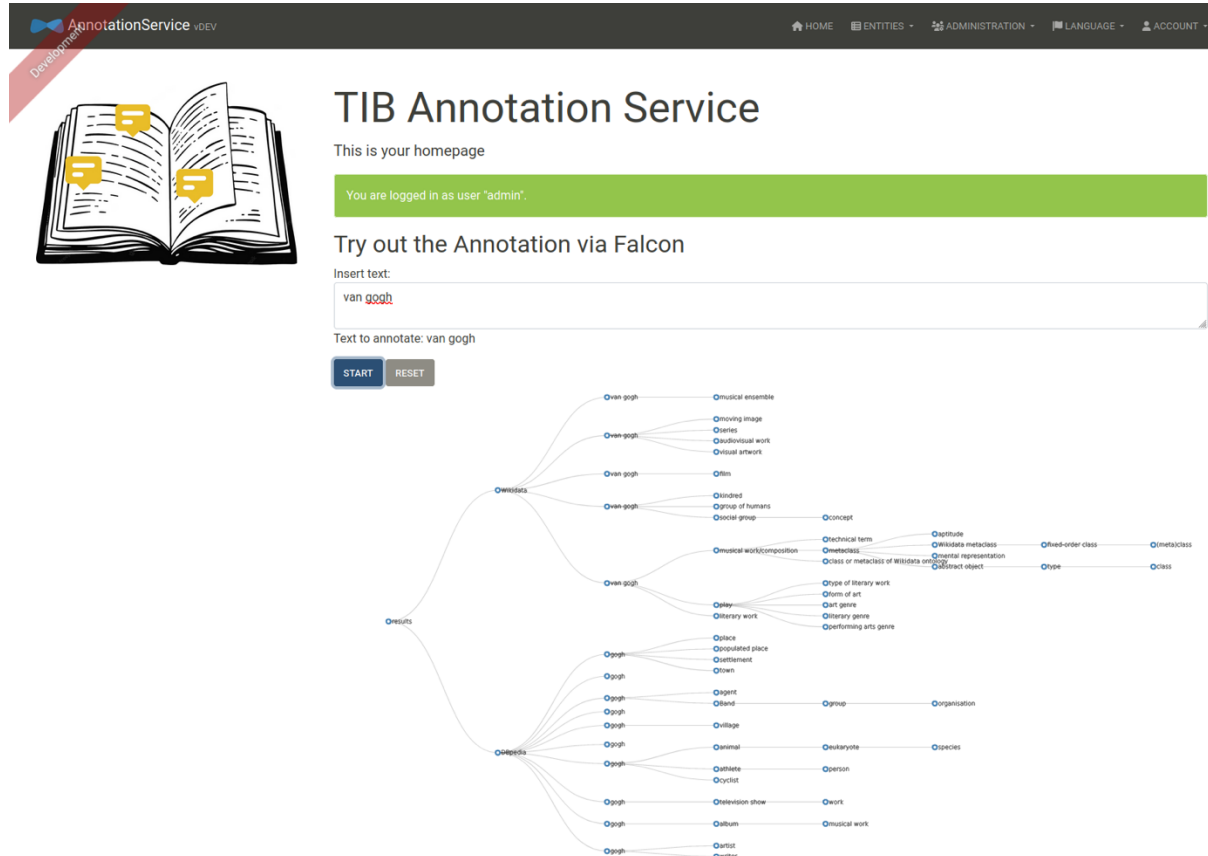
Some users also sketched out ideas for working with data dumps of existing terminologies and the possibility to query the dumps via an API.

Further important points noted during this exercise included the need to provide guidance to users when terminologies were in danger of getting deprecated, and the need to be able to search terminology services for terms in German directly, without need to translate to English first.

While these ideal user journeys pose multiple technical challenges, they also provide excellent guidance for the future development of new tools that address specific pain points and can be integrated into current workflows, instead of reinventing existing tools with only minor improvements.

Show and tell

After the practical exercises with participants. The Open Science Lab team presented early work on a new terminology and annotation service that aims to bridge the gap between existing services and collection management systems.

Screenshot image of the (work-in-progress) frontend interface of the annotation service web application showing search results visualisations from Wikidata and DBpedia. Kolja Bailly. CC-BY.

The team shared [a diagramme of the service architecture](); a short demo of a first prototype of the service, which relies on the open source software [Falcon]() to search terms in Wikidata and DBpedia and returns results with category trees visualised as network graphs; and a [short demo]() of the current OSL MVP (minimum viable product) integration between the media viewer Kompakkt and the LOD database software Wikibase – the last demo aimed to highlight the annotation capabilities of Kompakkt and how a custom integration with the annotation service API is planned to support semi-automated annotation workflows with standard terminologies. The team also shared the original [user stories]() drafted to support the design of the service. All these resources are also linked and available via the dedicated public [Gitlab repository]().

Outlook

The workshop concluded with a discussion reiterating the key requirements listed in the findings section of this report, and with a [call for feedback]() on future service developments at OSL. Upon requests from the participants to continue this line of open conversations, the workshop will likely turn into a series with future events providing further opportunities to test not only services developed at OSL, but other tools and terminology resources developed by members of the culture community.

To stay up-to-date with this workshop series and all other events planned by the NFDI4Culture LOD Working group, sign up to the [mailing list]() or [request to join]() the Rocket Chat channel.

Acknowledgements