

# Open Tool Registries!

Resolving the Directory Paradox with Wikidata

Till Grallert<sup>1,2,\*</sup> Sophie Eckenstaler<sup>1,3</sup>  
Claus-Michael Schlesinger<sup>1,3</sup> Nicole Dresselhaus<sup>1,2</sup>  
Isabell Trilling<sup>1,2</sup>

2025-03-20

## Abstract

This paper introduces the conceptual framework for open and community-curated tool registries, posing that such registries provide fundamental value to any field of research by acting as curated knowledge bases about a community’s past and current methodological practices as well as authority files for individual tools. This modular framework of a basic data model, SPARQL queries, bash scripts, and a prototypical web interface builds upon the well-established and open infrastructures of Wikimedia, GitLab, and Zenodo for creating, maintaining, sharing, curating, and archiving linked open data. We demonstrate the feasibility of this framework by introducing our concrete implementation of a tool registry for digital humanities, initially repurposing data from existing silos, such as TAPoR and the SSH Open Marketplace, and retaining the established TaDiRAH classification scheme while being open to communal editing in every aspect.

<sup>1</sup> Humboldt-Universität zu Berlin

<sup>2</sup> NFDI4Memory

<sup>3</sup> Kompetenzwerkstatt Digital Humanities

\* Correspondence: Till Grallert <till.grallert@hu-berlin.de>

## Introduction

### Tool Registries in the Digital Humanities

Registries collecting information on research tools, how they relate to methods, who already implemented them for their research questions, and how to learn to use and adapt them to your own research, address an obvious and concrete need of research communities and individual practitioners. Particularly in growing fields with rapidly evolving and complex technology stacks, such as the digital humanities, registries are envisioned to serve a central function within

the knowledge ecology, namely as one-stop shops for providing a comprehensive and up-to-date<sup>1</sup> overview of (computational) methods and their implementation in specific software. In practical terms, tool registries are of a dual nature. On the one hand, they embody a field’s methodological knowledge, the archive of scholarship as *techné*,<sup>2</sup> connecting information about tools to information about their use and purpose. On the other, registries depend on and form themselves part of the socio-technical infrastructures of the information age to maintain and access the record of that knowledge. It is evident that, at least in the practical reality of limited resources, their aims must remain aspirational in both their content and their infrastructural implementation.

Nevertheless, registries of computational tools have grown into a well-established genre within digital humanities (see Grant et al. 2020 for a history of tool registries in DH): from DiRT (*Directory of Research Tools*) to Bamboo and The Canadian *Text Analysis Portal for Research* (TAPoR 3), large EU projects like the *Social Sciences and Humanities Open Marketplace* (SSHOM) and DARIAH’s (*Digital Research Infrastructure for the Arts and Humanities*) now defunct *Tools E-Registry for E-Social science, Arts and Humanities* (TERESAH), or, in Germany, the consortia of the *National Research Data Infrastructure* (NFDI) and the Specialized Information Services (*Fachinformationsdienste*, FID), to individual libraries and institutions. Most tool registries approach knowledge about *techné* and the infrastructure needed to maintain this knowledge through an archival framework, with its fundamental hierarchy and power relations between knowledge production and knowledge consumption, housed and controlled by institutions unaccountable to the *démos* and enforced by the *archon* (chief magistrate) (c.f. Derrida [1995] 1996; Vismann 2009; Ebeling 2009). Albeit conceptually undemocratic, this is not to say, that the archive is governed by ill will. To the contrary. But academic knowledge ecologies are both inherently conservative and hierarchical. Tool registries seem a familiar task to which we can readily apply our well-established models and practices, from expert peer-review to curatorial hierarchies, to tightly controlled technical infrastructures of proprietary data silos, and grant-funded project cycles.

Quinn Dombrowski (2021) has summarised the fundamental shortcomings of existing tool registries against the backdrop of her own involvement with DiRT and the TADiRAH classification scheme, arguing that despite the considerable resources poured into each of them, they have failed as infrastructures and part of larger scholarly ecosystems (c.f. Bernardou et al. 2018). At best, they can, and probably should, be considered snapshots or documentations of historical practices at given points in time and space. This necessitates a shift in our perception of catalogues and registries. Users seek up-to-date information to

---

<sup>1</sup>c.f. Hughes, Constantopoulos, and Dallas (2015), 156 for topicality as a central defining quality of registries

<sup>2</sup>This, briefly, is inspired by Derrida’s ([1995] 1996); call for an *archivology* and his genealogy of the archive as grounded in the Latin *archivum* from Greek *arché* as *commencement* and *commandment*; Foucault’s ([1969] 2002, esp. 126-131) *Archaeology of Knowledge* and Heidegger’s writing on technology (Heidegger [1953] 2000; c.f. Ihde 2010, esp. 32-35, 62).

solve an issue at hand, while curators are painfully aware of any given catalogue entry’s quality as a historical artefact bound up in the contingencies of its making.

Gathering and curating information about a large and ever-growing number of tools requires substantial resources and skilled domain experts in a broad range of fields. The common solution to this problem is to build on existing bodies of knowledge or, in other words, to repackage existing information in new interfaces to address our patrons’ needs and our stakeholders’ interests—our libraries’ cataloguing ledgers have been superseded by index cards, which were then scanned into a (Card) Image Public Access Catalogue (CIPAC or IPAC), transcribed into digital texts, and fed into database systems that powered OPACs (Open Public Access Catalogue) (c.f. Oberhauser 2003), and which are now feeding large commercial aggregators. Regarding the genealogy of information on tools, take, for example, SSHOM. As part of the *European Open Science Cloud* (EOSC) under the European Commission’s *Horizon 2020* programme, SSHOM was built by DARIAH, the *Common Language Resources and Technology Infrastructure* (CLARIN), and the *Consortium of European Social Science Data Archives* (CESSDA) between 2019–23 and is considered a flagship service for DARIAH. In May 2023, SSHOM hosted information on almost 1700 items. More than 1200 of these were ingested from TAPoR over a few days in late 2022.<sup>3</sup> TAPoR in turn has a strong focus on textual research. Its current iteration integrated yet another popular tool registry, DiRT, in 2017/18, which itself had originated from earlier tool registries such as Bamboo (Grant et al. 2020; Dombrowski 2021, 2014; Rockwell 2006).

## A Modular and Open Tool Registry

The conceptual and technical approach to building the Wikidata-based, open tool registry described in this essay was developed against the backdrop of two interrelated use cases. Firstly, there was a need for a small and well curated list of tools to be used in a consulting and service context at the library (Grallert et al. 2024). Digital scholarship librarians or, more broadly, digital scholarship experts offering research support services, bring their own expertise regarding specific fields of research, scientific methods and digital tools. Library teams—including teams of one—integrate consulting services with other formats like workshops or even offer basic infrastructure as digital scholarship support and as technical partners for research projects. Institutionalized teams or larger institutional contexts often work with team-specific toolsets and technical stacks, which need to be documented for further reference, e.g. when documenting use cases or research output. Researchers, in turn, require information regarding the possibilities for implementing computational workflows adapted to specific

---

<sup>3</sup>This information is transparently provided on the item level through the API (see below). The website is much quieter about this fundamental heritage and only lists TAPoR 3 as one of 15 “trusted sources” (“About the Data Population” 2023).

research questions and research data – from aggregation to publication. In such contexts, curated lists of tools reflect team expertise and recommendations for tools optimized for local target groups.

Secondly, we were looking for a publicly-accessible authority file of tools, a single source-of-truth about software, which would allow us to unambiguously identify and reference tools employed in the course of a research project. Such an authority file would also greatly facilitate the ability to quickly and transparently curate subsets without a need for manually maintaining and updating redundant information. A catalogue to follow up on methods and tools mentioned in other scholars’ research output.

The approach proposed here is designed to address both use cases by providing the basic infrastructure to put together a well-curated data set for local communities and making it possible to store aggregated and enhanced datasets ready for reuse.

The design goals for the system presented here are oriented towards the target groups of librarians, researchers, and institutions. As all of these groups provide valuable information on tools and methods and, at the same time, might make good use of descriptions and documentation provided by peers, all components of the system should be open, sustainable, re-usable and adaptable. All components for local deployment and components for necessary documentation and discussion, e.g. regarding the basic data model, should meet these criteria. Data produced through application of modules or components should meet the FAIR criteria (“FAIR Principles” 2020), and in addition should be browsable and comparable using elements provided through workflow description or basic modules, e.g. a web frontend with a search interface and browsing functions.

While platforms and web applications tend to pull users to their specific solution, a lot of data aggregated by larger tool registry projects is readily available as structured data. This makes it possible to build and curate subsets based on data points and corresponding queries and, from there, offers the possibility for an open and modular approach, with modules consisting of a simplified data model, a data store, data curation functions and a web frontend for data presentation.

Our set-up is inspired by *minimal computing* (Gil and Ortega 2016; Risam and Gil 2022) and the *Endings Principles for Digital Longevity* (Endings Project Team 2023). Optimizing for sustainability and adaptive reuse, we chose a consequent separation of modules, minimized complexity and opted for openness on every level. Making use of existing infrastructure whenever possible minimizes overhead for hardware and software. Using well-known standard solutions opens up possibilities for interacting with the data (sustain and develop) and making use of the data (search and analyse). We use Wikidata as a data store, which allows for open access and positions the data as part of the commons, and, in a best case scenario, the tool registry data as the object of a collaborative effort to aggregate, sustain and develop information in regard to the DH

community and, furthermore, also other communities working with structured descriptions of software, e.g. due to the growing necessity for process metadata or software preservation purposes (Christophersen et al. 2023). In order to address the problem of a back-end infrastructure outside of our control and a scholarly-wariness of vandalism, we archive weekly, versioned snapshots of the dataset via the publicly-funded Zenodo repository for research data (for more details see the section on workflows below) (Grallert and Dresselhaus 2024).

We provide a basic object-oriented data model for tools as objects as a reference model. We consider this reference model as a minimal basic set that can and should be extended with domain specific and research specific elements. This means that while everyone is able to use their own models, as long as these models respect the basic reference model, all information added to Wikidata will immediately become part of the shared tool registry. A community-oriented documentation and discussion process based on the Wikiproject documentation provides a channel for working on the common reference and evolving the data model regarding mandatory and optional qualities as well as additional modules developed for specific purposes (“Wikidata:WikiProject DH Tool Registry” 2024).

Wikidata provides several interfaces, e.g. a web frontend, a web API and a SPARQL endpoint. As all of these interfaces come with a certain amount of complexity that necessitate prior knowledge about Wikidata data structure and principles, a web interface module provides functions for curating, exploring and displaying datasets based on the Wikidata pool. The web interface is based on a template which can be turned into a static HTML site using a static site generator, optimizing for sustainability and adaptability.

## Modular Application Structure

### Data models

Wikidata follows no strict ontology but instead applies an open, bottom-up approach, which is famously referred to as a *folksonomy* (a term coined by VanderWal (2004)), with a very basic and deliberately generic data model of *items* (identifiers starting with “Q”) and *properties* (identifiers starting with “P”) connecting items to other items or *values* of various data types (strings, dates, integers ...) through *statements*. All statements can carry properties as qualifiers and references to the source for this information. A large part of the statements in Wikidata are properties linking to other platforms and authority files through external identifiers. All properties might come with a set of formal constraints, which, if violated, will result in various warning flags in the Wikidata web interface. Finally, all items should carry two special string-value properties for multilingual human-readable labels and descriptions. The only mandatory property is instance of (P31) linking to one or more items, without which an item would not be connected to the larger knowledge graph. But even this constraint

is not strictly enforced on a technical level (c.f. Hosseini Beghaeiraveri et al. 2023a).

Such a generic and unconstrained data model makes it hard to query Wikidata with a predictive result but it provides a system for any conceivable, community-driven data modelling: Data models exist largely through communities of practice adhering to them and providing the necessary SPARQL queries for building sub graphs. To this end we set up a WikiProject as a port of entry to build a community of people and institutions wanting to maintain a tool registry. This WikiProject documents our approach and data model (“Wikidata:WikiProject DH Tool Registry” 2024). As an added bonus, WikiProjects rank extremely high on search engines.

### Our basic data model

We differentiate between a number of concepts to model the relation between a tool and its purpose:

1. research **tools** comprise both **methods** and concrete **software**
2. **methods** are informed by **theories** and have a *purpose*.
3. **methods** are implemented through (multiple) layers of **software**, which, in turn, require *hardware* and infrastructural resources such as electricity, internet connectivity or licences and which interact with data *formats* and *serialisations* (reading and writing).
4. **software** is written in programming languages and can be interacted with through *interfaces*. Command line interfaces (CLI), including application programming interfaces, require knowledge of programming languages to interact with them.
5. **methods**, **languages**, and **formats** rely on and implement abstract *concepts*

The tool registry is concerned with only a subset of this larger ontology: *software* and *methods*. The basic data model for *software* requires no mandatory information in addition to the Wikidata base model of *label*, *description*, and instance of (P31) beyond the has use (P366) property that associates *software* to one or more *methods*.

The basic data model for *methods* is even more rudimentary. Here the only mandatory property is a TaDiRAH ID (P9309) that proclaims equivalence to a concept within the “The Taxonomy of Digital Research Activities in the Humanities” (TADiRAH, see below). fig. 1 shows a schematic of this data model conceptualising “Gephi” as an instance of “Software” that can be used for “network analysis”. Additionally, the source for this claim is identified through a reference to the URI of an entry in the SSHOM.

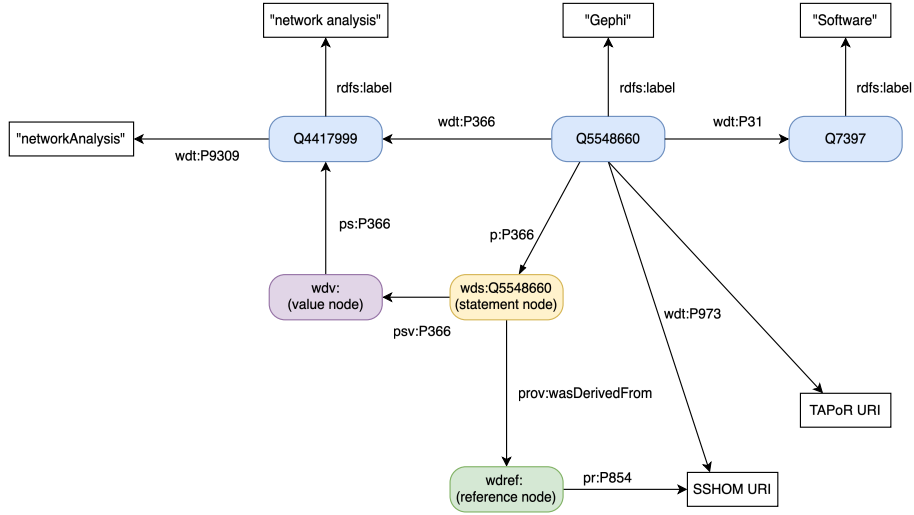


Figure 1: Schematic of our basic data model, using *Gephi* (Q5548660) as an example

### TaDiRAH mapping

Like all archives, tool registries depend on classification schemes and taxonomies in order to file, retrieve, and produce knowledge about an item in their collection. Over the course of the last decade “The Taxonomy of Digital Research Activities in the Humanities” (TADiRAH) has become the most widely adopted classification scheme in the digital humanities and is used in its current version (>2.0) for author-assigned classifications of conference submissions to the *Alliance of Digital Humanities Organizations* (ADHO) and *Digital Humanities im deutschsprachigen Raum* (DHD) conferences as well as SSHOM (TAPoR uses the older, incompatible version of TADiRAH).

TADiRAH was developed from the mid-2010s onwards in the tradition of John Unsworth’s “scholarly primitives” (For the genealogy of TADiRAH see Borek et al. 2016; Hughes, Constantopoulos, and Dallas 2015, 155–57). Members of DARIAH-DE and BambooDiRT developed TADiRAH on the basis of the *ICT Methodology*, which itself had evolved from the AHDS Taxonomy of Computational Methods (2003–) under the auspices of the Oxford University Digital Humanities Programme. The first version of TADiRAH (v0.5) was released in 2016 (Borek et al. 2016; Borek et al. [2014] 2015) as a four-level hierarchy of *goals*, *methods*, *techniques*, and *objects*. The current iteration (>v2.0) was released in 2021 (Borek et al. 2021) and has been redesigned from the ground up as a SKOS vocabulary (Simple Knowledge Organization System). Unlike earlier iterations, it only covers methods and techniques as a single type of entities, which can nest in three hierarchical layers. Therefore, v2.0 is not backwards compatible with earlier versions.

The major strengths of TADiRAH are its wide adoption across the field of digital humanities and its integration into wider Linked Open Data (LOD) infrastructures as part of the DARIAH Vocab services, which is maintained by the *Austrian Centre for Digital Humanities and Cultural Heritage* (ACDH-CH) of the *Austrian Academy of Sciences*. In addition, a partial mapping between TADiRAH and Wikidata has been implemented on both sides. The Wikidata property TaDiRAH ID (P9309) was created in 2015 by Adam Schiff, Principal Cataloger at the University of Washington Libraries in Seattle, to state equivalence between a Wikidata item and a TADiRAH class.

However, TADiRAH has major flaws rather similar to those of tool registries:

1. Development depends on voluntary labour and thus the vocabulary has been dormant for a number of years.
2. The current version and its documentation are return v0.5 as it is hosted on GitHub (see also Zhao 2022).
3. The SPARQL endpoint and API, which would serve the classification scheme as RDF, are frequently down.

Despite of these weaknesses, potential competitors and successors have not been successful. The “NeDiMAH Methods Ontology for the Digital Humanities” (NEMO) (Hughes, Constantopoulos, and Dallas 2015, 165–67; Bernardou et al. 2018, 4–5), developed with substantial funding from the European Science Foundation by the *Network for Digital Methods in the Arts and Humanities* (NEDiMAH) from 2011 to 2015 (Hughes, Constantopoulos, and Dallas 2015) is a CIDOC CRM-compliant ontology and provided a mapping from TADiRAH but has seen no tracable adoption.<sup>4</sup> This might be due to information on NEMO—quite fittingly, given its name—being incredibly hard to come by. The official website is outdated and has not been updated with the project results and access to the documentation, although officially licensed under “CC BY-NC-SA 4.0”, requires a login on <http://nemo.dcu.gr/resources/>.

Given its wide adoption, including our seeding data set of tools from SSHOM and TAPoR, we have therefore opted to retain TADiRAH classification as the main organising principle for our tool registry. The system is flexible enough to adopt other classification systems should they emerge, as long as their identifiers are mapped to a Wikidata property.

The TADiRAH classification of items is implemented through the *has use* property, linking to other Wikidata items that carry a TaDiRAH ID (P9309). To this end we completed the mapping between Wikidata and TADiRAH so that all TADiRAH classes can be found on Wikidata (see below).

---

<sup>4</sup>The academic knowledge graph OpenAlex finds only 9 references.



## Optional parts of the basic data model and curating collections of tools

The basic data model leverages the weaknesses of an open-world knowledge graph without a formal ontology to our advantage as it can be easily extended to accommodate additional information on individual items, such as licenses, version numbers, URIs of a source code repository etc. For this purpose, our basic data model proposes a number of optional statements.

A core requirement for our tool registry is the ability for diverse communities to curate their own collections. This can easily be done through the property *collection* (P195) pointing to one or more Wikidata items describing the collection and potentially naming curators and contributors (fig. 2). We have implemented two such collections, *SSHOM* (Q131847864) and *TAPoR* (Q3979414), to document existing tool registries that formed the basis of our dataset (see below). Note that such collections are not necessarily limited to tools classified with TADIRAH.

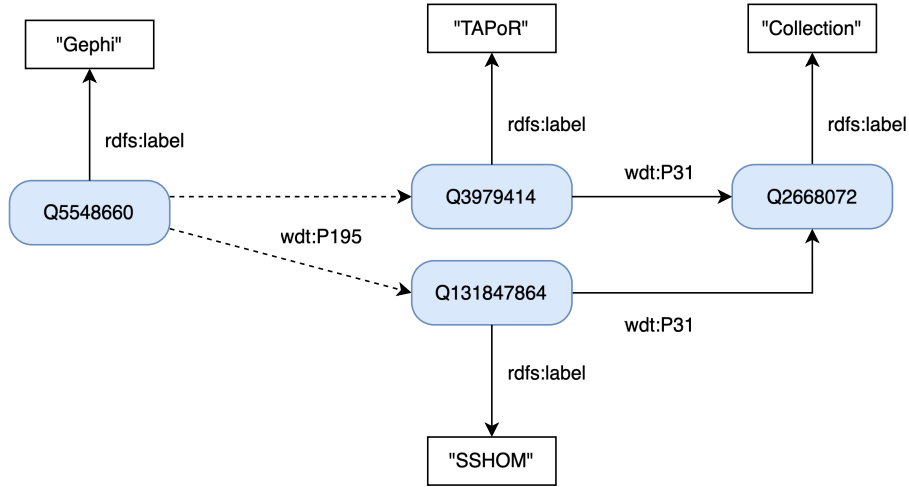


Figure 2: Schematic of the data model for modelling registries as collections, using *Gephi* (Q5548660) as an example

## Domain-specific extensions

This approach also allows to accommodate the needs of other communities that might want to model additional, domain-specific relations. One such extension has been implemented by some of the authors as part of a survey of the fields of digital humanities and digital history. There we model the relation between methods (classified with TADIRAH), research output in the form of publications and conference papers, and their authors. This extension then allows to ask for exemplary applications of a method and the tools (potentially) used by specific scholars (Grallert, Trilling, and Skibba 2025), which, in turn, allows to track the

popularity of tools over time or to build training curricula based on the relevance of particular tools for specific fields (see for instance `method_authors.rq` from Grallert (2024)).

## Frontends

Wikidata’s default interface emphasises its use as an authority file by focusing on individual items and provides all available information on an item via the so-called *Linked Data Interface* fig. 3. The Linked Data Interface functions both as a view and as an input mask, enabling rather efficient data processing for a graphic user interface. This is greatly facilitated by the simple presentation of a list of statements currently available for any specific item, including all qualifiers and a summary of existing references. However, this reduced presentation has its own disadvantages. For example, the rather technical design of the interface, which is based on the principles of the Linked Open Data (LOD) concept (Berners-Lee 2009), can be perceived as cumbersome and less intuitive, especially for new users and those unfamiliar with knowledge graphs.

But most importantly, by focusing on individual items, the interface does not provide direct access to the underlying knowledge graph. This also means that there is no obvious or simple means for accessing specific subsets of the knowledge graph, which is a particular challenge for use in the context of the specialised tool registry.<sup>5</sup> For this purpose, Wikidata provides a powerful tool outside the Linked Data Interface: the SPARQL Query Service Tool. However, this requires a profound knowledge of the SPARQL query language, which makes it difficult to use for many potential users.

## Query tools with SPARQL

As stated above, the data model is fundamentally instantiated through SPARQL queries. We therefore provide a set of modular queries for use in various applications and to ease access to SPARQL for colleagues less comfortable writing their own queries from scratch (Grallert 2024). Each query first asks for all items with a TADIRAH ID, that is items, which our model considers to be *methods*. It then queries for all items, which point to at least one of those methods through the has use property and which are instances of “Software” or its subclasses:

```
SELECT DISTINCT
  ?tool ?toolLabel # only get Software-ID and Software-Name
WHERE {
  ?method wdt:P9309 ?tadirahID.           # Variable method is a tadirah-method
```

---

<sup>5</sup>Hosseini Beghaeiraveri et al. (2023b) showed that subsetting massive knowledge graphs such as Wikidata has become more relevant in recent years, especially in the scientific context with regard to reuse and archiving. However, the selection of approaches and tools evaluated in the study also makes it clear that there is still no standardised procedure that can be easily applied to the tool registry.

## OCR4all (Q124347709)

An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings

 [edit](#)

[In more languages](#)

[Configure](#)

Language	Label	Description	Also known as
English	OCR4all	An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings	
German	No label defined	No description defined	
French	No label defined	No description defined	
Bavarian	No label defined	No description defined	

[All entered languages](#)

### Statements









instance of	 software	 <a href="#">edit</a>
	<a href="#">▼ 0 references</a>	<a href="#">+ add reference</a>
	 research tool	 <a href="#">edit</a>
	<a href="#">▼ 0 references</a>	<a href="#">+ add reference</a>
	 research software	 <a href="#">edit</a>
	<a href="#">▼ 0 references</a>	<a href="#">+ add reference</a>
		<a href="#">+ add value</a>
has use	 optical character recognition	 <a href="#">edit</a>
	<a href="#">▼ 0 references</a>	

Figure 3: The item *OCR4all* (Q124347709) in the Linked Data Interface of Wikidata.

```

?tool wdt:P366 ?method;          # Variable tool 'has method' method
      (wdt:P31/(wdt:P279*)) wd:Q7397. # and tool is child of "Software"
SERVICE wikibase:label {
  # set wikibase-service to auto-language with fallback english
  bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".
  # get the tool-label (=name) of our tool.
  ?tool rdfs:label ?toolLabel.
}
}

```

This query returns the IDs of items and their label as a potential sanity checker for human readers. The ID can then be used in further SPARQL queries, API calls, or the plethora of tools for interacting with Wikidata hosted on Toolforge, ranging from *Scholia*, a long-running project for querying and visualising scientometrics (Nielsen, Mietchen, and Willighagen 2017), to *Reasonator*, which displays Wikidata items in a view optimised for the item-type and enhanced with some basic reasoning.

Another option is to directly query for tools in curated collections, such as SSHOM (Q131847864) and TAPoR (Q3979414), as mentioned above.

```

#title:Tools in the SSHOM
#defaultView:Table
PREFIX collection: <http://www.wikidata.org/entity/Q131847864> # a specific collection
SELECT
  ?tool ?toolLabel
WHERE {
  ?tool wdt:P195 collection: ;      # items in the collection
        wdt:P31/wdt:P279* wd:Q7397. # limit tools to software in the broadest sense
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".
    ?tool rdfs:label ?toolLabel.
  }
}
LIMIT 3000

```

## Tool Registry Frontend Module

In order to overcome these limitations and meet the specific requirements of the tool registry for use in the context of the *Kompetenzwerkstatt Digital Humanities* (KDH) at the university library of the Humboldt-Universität zu Berlin, we developed a separate frontend. The idea of using Wikidata as a database and authority file for specific web applications has already been implemented by other projects. Good examples for this approach are the *Archivführer Kolonialzeit* and *Scholia*. The core argument for this architecture is that thanks to Wikidata APIs one can build custom interfaces for specific communities while retaining all the advantages of Wikidata communities of contributors and its

centralised and persistent data storage without being locked into its user interface.

Our front end offers a specialised search for digital humanities tools, as well as a more reader-friendly individual view of the tools compared to the linked data interface. It also enables the integration of corporate design elements and the provision of additional background information relevant to the KDH project. Care was taken to strictly maintain the separation between data and presentation levels in order to ensure data integrity and reusability.

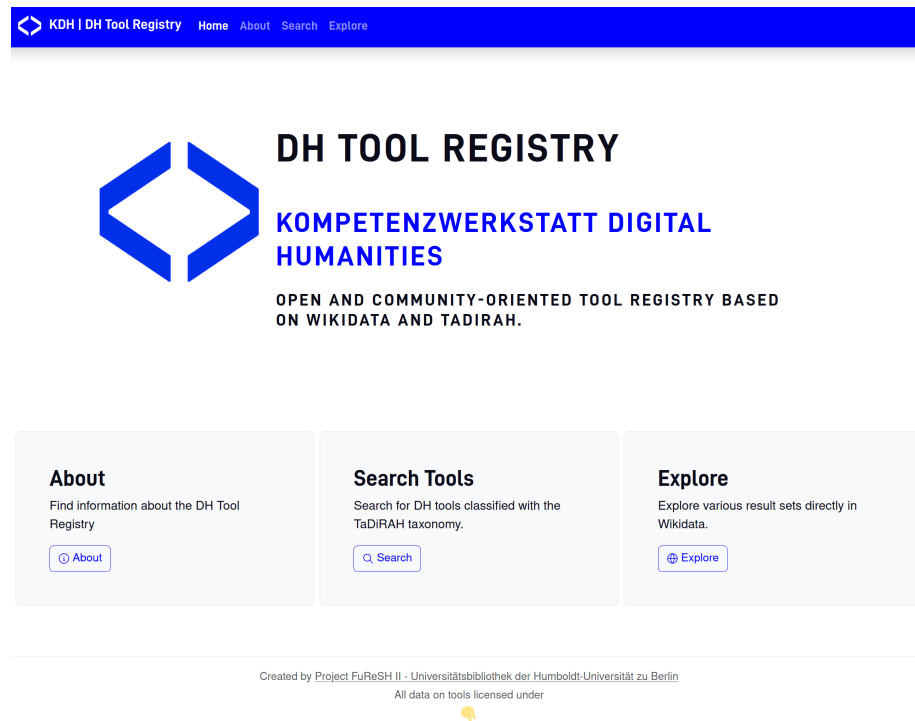


Figure 4: The front-end application developed in the KDH uses Wikidata as a database and authority file.

In order to facilitate installation and operation and to minimise the hurdles for subsequent use by other projects, the frontend was implemented as a static website and published via GitHub pages.<sup>6</sup> This decision minimises the maintenance effort and makes it easy to deploy the frontend.

<sup>6</sup>The source code of the front-end application is also published on GitHub: <https://github.com/FuReSH/tool-storage-interface>.

## Data and Workflows

### Data sources

As outlined above, tool registers tend to build upon existing data sets to save time and effort on curation. We follow this approach to populate our tool registry with the existing knowledge about tools within the DH community by integrating and linking the TAPoR and SSHOM data sets and the classification of tools with TADiRAH into Wikidata. All three platforms provide machine-actionable data through APIs and serialised as JSON.

Unfortunately, TAPoR’s APIs are entirely undocumented and we only discovered them through other projects interested in the usage of tools within the DH community, namely the ToolXtractor (Fischer, yoannspace, and laureD19 [2019] 2022; see also Barbot et al. 2019a, 2019b; Fischer and Moranville 2020). The TAPoR API provides at least two endpoints. [https://tapor.ca/api/tools/by\\_analysis](https://tapor.ca/api/tools/by_analysis) returns a full list of all tools with minimal information for each entry. Importantly, it links tools to categories of research activities, e.g. *Open-Refine* to “Enrichment”. <https://tapor.ca/api/tools/%7BBID%7D> then returns detailed information for individual tools, based on the tool ID obtained through the first endpoint.<sup>7</sup>

SSHOM’s API, on the other hand, is well documented. Again, one has to query for all tools (through <https://marketplace-api.sshopencloud.eu/api/tools-services?approved=true>) in order to retrieve the necessary IDs for individual tools, which then allow to use more specific API endpoints such as <https://marketplace-api.sshopencloud.eu/api/tools-services/%7BBID%7D>. In either case the API returns plenty of data, including links to equivalent concepts on Wikidata and classifications such as TADiRAH or the *Austrian Fields of Science and Technology Classification 2012* on the DARIAH Vocab services. The APIs, however, do not allow to query for tools by concepts from these classification schemes. It seems that this is only possible through the web interface and URL query parameters, i.e. <https://marketplace.sshopencloud.eu/search?f.activity=%7Bactivity%7D> where the camel case in TADiRAH concepts is replaced with “+” and all component terms are capitalised, i.e. `opticalCharacterRecognition` needs to be translated to `Optical+Character+Recognition`.

TADiRAH is a SKOS vocabulary currently hosted through the DARIAH Vocab services. Even though this platform is envisioned as part of the Semantic Web and should provide Linked Open Data, it has a patchy track record of doing so. At the time of writing in 2024, neither the service’s API and SPARQL endpoint nor their documentation can be reached. Fortunately, we were able to download all data as RDF in late 2022.

---

<sup>7</sup>Information includes data on tools such as URLs, source code repositories, email addresses of creators, image URLs, metadata on when the TAPoR entry was last updated and by whom, as well as rating on the platform

## Data import

There are generally two main methods for adding and editing Wikidata items in bulk: QuickStatements and OpenRefine. We chose the latter as we had to reconcile our initial datasets with existing items to avoid creating duplicates. In addition, OpenRefine provides means for extensive data manipulation and a graphical user interface.

Reconciliation required a lot of semi-automated data cleaning and manual decision making. The necessary schemas for mapping our input data sets to the data model were created iteratively through the OpenRefine GUI and in tandem with the data model. In total, we have mapped 85 missing TADIRAH classes (mostly through linking TADIRAH IDs to existing Wikidata items), added almost 700, and edited more than 1200 tools from the SSHOM and TAPoR data sets.<sup>8</sup>

Wikibase, the software running Wikidata, provides a mature system of user management and version control, which allows to revert to earlier states. There are some tools for editing and reverting batch edits, but at the time of writing in 2024 they had become dysfunctional and were looking for new maintainers.

## Adding and editing data

Wikidata allows anyone with an internet connection to edit and contribute without further ado. Registration is not required, but the platform will in such cases log the editor’s public IP address. With our approach, existing queries will pick up new information as soon as they are added to Wikidata. Tracking changes to “our” data set, we see constant improvements of data through a combination of individual edits through the Wikidata user interface, bulk edits, and bots. The latter, for instance, periodically query linked GitHub repositories for new releases and update items accordingly.

## Export and publication of stable data sets

One of the primary challenges in working with Wikidata in academic environments lies in its open-editing nature; anyone can modify data entries, which might raise concerns about reliability and accuracy. This characteristic necessitates additional measures to ensure the stability and trustworthiness of the information it contains, especially for academic use where data integrity is paramount for reproducibility.

To address this issue, we regularly pull copies of our dataset from Wikidata in formats such as RDF (serialised as Turtle, `.ttl`) or JSON-LD using a combination of SPARQL queries and API calls. We publish these versioned releases on platforms like GitLab or GitHub and archive them on public repositories such as Zenodo (Grallert and Dresselhaus 2024).

---

<sup>8</sup>The JSON schemas for uploads from OpenRefine to Wikidata can be found at [https://scm.cms.hu-berlin.de/methodenlabor/p\\_publish2wikidata](https://scm.cms.hu-berlin.de/methodenlabor/p_publish2wikidata).

This approach allows the datasets to be cited with a DOI, ensuring that specific versions can be referenced reliably. Wikidata itself provides persistent identifiers for citing particular versions of data objects, but through the Linked Data Interface the statements of such persistent versions of items still point to the canonical ID of other items. Similarly our dataset currently only includes full copies of first-level objects (tools) and labels and descriptions for all linked items. Full access to historical perspectives on the dataset would require to archive the entire subgraph with all dependencies that yields a great amount of Wikidata itself (c.f. Hosseini Beghaeiraveri et al. 2023a) and is not suitable with platforms like Zenodo.

## Discussion and outlook

In this paper we have presented a conceptual framework and concrete implementation of what we believe to be a sustainable approach to the directory paradox as formulated by Dombrowski (2021). Our Wikidata-based tool registry for digital humanities attends to communities, who need a flexible and open authority file, as well as to those, who are interested in linked open data, allowing them to curate their own subsets or to query the graph for connections beyond the initial scope of a tool registry. We are fully aware that the fundamental flexibility and openness of Wikidata and the absence of hierarchical access control and formal ontologies, such as CIDOC-CRM, will impede adoption among libraries and infrastructure providers. However, the QIDs of tools can simply function as a persistent identifier (PID) to link otherwise disparate registries and knowledge graphs together.

While TADIRAH is currently the only classification scheme for research in the digital humanities employed across multiple venues, it can be superseded or amended by future classification schemes for the digital humanities when they emerge. Equally, other fields could use their own classification schemes without breaking the data model and registry introduced in this paper by adapting and modifying the relevant SPARQL queries.

We acknowledge that the reliance on SPARQL is a major obstacle to wider adoption. We thus perceive of the modular frontend and a graphical user interface for building queries as the most important venues of future development. We can also imagine a frontend with editing functionality based on HTML forms, JSON schemas, and the Wikidata REST API currently under development (“Starting Fresh: The Wikibase REST API” 2023) in order to technically enforce data models.



## Acknowledgments

### Funding

This work was funded by the German Research Foundation (DFG) through a collaboration between the NFDI consortium 4Memory (www.4memory.de, DFG project no. 501609550) and *Future e-Research Support in the Humanities II* (FuReSH II, DFG project no. 466522693).

### Author contributions (CRedit)

- Conceptualization: Till Grallert, Sophie Eckenstaler, Claus-Michael Schlesinger
- Data Curation: Till Grallert, Isabell Trilling
- Software: Sophie Eckenstaler (Frontend, SPARQL), Nicole Dresselhaus (Archiving), Till Grallert (SPARQL, R)
- Writing – original draft: Till Grallert, Claus-Michael Schlesinger, Sophie Eckenstaler, Nicole Dresselhaus
- Writing – review & editing: Till Grallert, Claus-Michael Schlesinger

### Data availability

All data and code is available on Zenodo:

- SPARQL queries: Grallert (2024).
- Data model and JSON schemas for use with OpenRefine: [https://scm.cms.huberlin.de/methodenlabor/p\\_publish2wikidata](https://scm.cms.huberlin.de/methodenlabor/p_publish2wikidata).
- Front end: <https://github.com/FuReSH/tool-storage-interface>.
- Weekly screenshots of the data set as exported from Wikidata: Grallert and Dresselhaus (2024).

## Bibliography

- “About the Data Population.” 2023. Social Sciences & Humanities Open Marketplace. August 31, 2023. <https://marketplace.sshopencloud.eu/about/data-population>.
- Barbot, Laure, Frank Fischer, Yoann Moranville, and Ivan Pozdniakov. 2019a. “Tools Mentioned in the Proceedings of the Annual ADHO Conferences (2015–2019).” December 6, 2019. <https://lehkost.github.io/tools-dh-proceedings/index.html>.
- . 2019b. “Which DH Tools Are Actually Used in Research?” *weltliteratur.net: A Black Market for the Digital Humanities*. December 6, 2019. <https://weltliteratur.net/dh-tools-used-in-research/>.
- Bernardou, Agiatis, Eric Champion, Costis Dallas, and Lorna M. Hughes. 2018. “Introduction: A Critique of Digital Practices and Research Infrastructures.” In *Cultural Heritage Infrastructures in Digital Humanities*, edited by Agiatis

- Bernardou, Eric Champion, Costis Dallas, and Lorna M. Hughes. Digital Research in the Arts and Humanities. London: Routledge.
- Berners-Lee, Tim. 2009. "Linked Data - Design Issues." June 18, 2009. <https://www.w3.org/DesignIssues/LinkedData#fivestar>.
- Borek, Luise, Quinn Dombrowski, Jody Perkins, and Christof Schöch. 2016. "TaDiRAH: A Case Study in Pragmatic Classification." *Digital Humanities Quarterly* 10 (1). <http://www.digitalhumanities.org/dhq/vol/10/1/000235/000235.html>.
- Borek, Luise, Quinn Dombrowski, Jody Perkins, Christof Schöch, and Matthew Munson. (2014) 2015. "TaDiRAH." Digital Humanities Taxonomy Group. <https://github.com/dhtaxonomy/TaDiRAH>.
- Borek, Luise, Canan Hastik, Vera Khramova, Klaus Illmayer, and Jonathan D. Geiger. 2021. "Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR." In *Information Between Data and Knowledge*, 321–32. Schriften Zur Informationswissenschaft 74. Glückstadt: Werner Hülsbusch. <https://doi.org/doi.org/10.5283/epub.44951>.
- Christophersen, Allan, Colón-Marrero Elena, Dianne Dietrich, Patricia Falcao, Claire Fox, Karen Hanson, Allen Kwan, and Matthew McEniry. 2023. "Software Metadata Recommended Format Guide." Zenodo. <https://doi.org/10.5281/zenodo.10001787>.
- Derrida, Jacques. (1995) 1996. *Archive Fever: A Freudian Impression* [Mal d'archive]. Translated by Eric Prenowitz. Chicago: University of Chicago Press.
- Dombrowski, Quinn. 2014. "What Ever Happened to Project Bamboo?" *Literary and Linguistic Computing* 29 (3): 326–39. <https://doi.org/10.1093/lc/fqu026>.
- . 2021. "The Directory Paradox." In *People, Practice, Power: Digital Humanities Outside the Center*, edited by Anne B. McGrail, Angel David Nieves, and Siobhan Senior. Debates in the Digital Humanities. Minneapolis: University of Minnesota Press. <https://dhdebates.gc.cuny.edu/read/people-practice-power/section/ca87ec4c-23a0-452d-8595-7cfd7e8d6f0c#ch06>.
- Ebeling, Knut. 2009. "Das Gesetz des Archivs." In *Archivologie: Theorien des Archivs in Wissenschaft, Medien und Künsten*, edited by Markus Knut Ebeling, Stephan Günzel, and Aleida Assmann, 61–88. Berlin: Kulturverlag Kadmos.
- Endings Project Team. 2023. "Endings Principles for Digital Longevity." <https://endings.uvic.ca/principles.html>.
- "FAIR Principles." 2020. GO FAIR. August 5, 2020. <https://www.go-fair.org/fair-principles/>.
- Fischer, Frank, and Yoann Moranville. 2020. "Tools Mentioned in DH2020 Abstracts." [weltliteratur.net](http://weltliteratur.net): A Black Market for the Digital Humanities. July 23, 2020. <https://weltliteratur.net/tools-mentioned-in-dh2020-abstracts/>.
- Fischer, Frank, yoannspace, and laured19. (2019) 2022. "ToolXtractor." <https://github.com/lehkost/ToolXtractor>.
- Foucault, Michel. (1969) 2002. *The Archaeology of Knowledge* [L'Archéologie Du Savoir]. Translated by A. M. Sheridan Smith. London: Routledge.

- Gil, Alex, and Élika Ortega. 2016. “Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing.” In *Doing Digital Humanities: Practice, Training, Research*, edited by Constance Crompton, Richard J Lane, and Ray Siemens, 22–34. Abingdon: Routledge.
- Grallert, Till. 2024. “Tool Registry: SPARQL Queries.” [https://scm.cms.huberlin.de/methodenlabor/tr\\_sparql](https://scm.cms.huberlin.de/methodenlabor/tr_sparql).
- Grallert, Till, and Nicole Dresselhaus. 2024. “Tool Registry for Digital Humanities.” Zenodo. <https://doi.org/10.5281/zenodo.14259807>.
- Grallert, Till, Sophie Eckenstaler, Samantha Tirtohusodo, and Claus-Michael Schlesinger. 2024. “Ob Werkzeugkoffer, Werkstatt Oder Baumarkt: Offene, Community-Kuratierte Tool Registries Mit Wikidata.” Poster. Zenodo. <https://doi.org/10.5281/zenodo.10698252>.
- Grallert, Till, Isabell Trilling, and Anica Skibba. 2025. “Wikidata:WikiProject Field Survey Digital Humanities / Digital History.” Wikidata. January 8, 2025. [https://www.wikidata.org/w/index.php?title=Wikidata:WikiProject\\_Field\\_Survey\\_Digital\\_Humanities/\\_Digital\\_History&oldid=2296127677](https://www.wikidata.org/w/index.php?title=Wikidata:WikiProject_Field_Survey_Digital_Humanities/_Digital_History&oldid=2296127677).
- Grant, Kaitlyn, Quinn Dombrowski, Kamal Ranaweera, Omar Rodriguez-Arenas, Stéfan Sinclair, and Geoffrey Rockwell. 2020. “Absorbing DiRT: Tool Directories in the Digital Age.” *Digital Studies / Le Champ Numérique* 10 (1). <https://doi.org/10.16995/dscn.325>.
- Heidegger, Martin. (1953) 2000. “Die Frage nach der Technik.” In *Vorträge und Aufsätze*, edited by Friedrich-Wilhelm von Herrmann, 7:5–36. Gesamtausgabe. Frankfurt: Klostermann.
- Hosseini Beghaeiraveri, Seyed Amir, Jose Emilio Labra Gayo, Andra Waagmeester, Ammar Ammar, Carolina Gonzalez, Denise Slenter, Sabah Ul-Hasan, Egon Willighagen, Fiona McNeill, and Alasdair J. G. Gray. 2023a. “Wikidata Subsetting: Approaches, Tools, and Evaluation.” *Semantic Web* Preprint (January): 1–27. <https://doi.org/10.3233/SW-233491>.
- . 2023b. “Wikidata Subsetting: Approaches, Tools, and Evaluation.” Edited by Lucie-Aimée Kaffee, Simon Razniewski, and Pavlos Vougiouklis. *Semantic Web*, December, 1–27. <https://doi.org/10.3233/SW-233491>.
- Hughes, Lorna, Panos Constantopoulos, and Costis Dallas. 2015. “Digital Methods in the Humanities: Understanding and Describing Their Use Across the Disciplines.” In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 150–70. Chichester: Wiley. <https://doi.org/10.1002/9781118680605.ch11>.
- Ihde, Don. 2010. *Heidegger’s Technologies: Postphenomenological Perspectives*. Perspectives in Continental Philosophy. New York: Fordham University Press.
- Nielsen, Finn Årup, Daniel Mietchen, and Egon Willighagen. 2017. “Scholia, Scientometrics and Wikidata.” In *The Semantic Web: ESWC 2017 Satellite Events*, edited by Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig, 237–59. Lecture Notes in Computer Science. Cham: Springer International Publishing. [https://doi.org/10.1007/978-3-319-70407-4\\_36](https://doi.org/10.1007/978-3-319-70407-4_36).

- Oberhauser, Otto Carl. 2003. “Card-Image Public Access Catalogues (CIPACs): An International Survey.” *Program: Electronic Library and Information Systems* 37 (2): 73–84. <https://doi.org/10.1108/00330330310472867>.
- Risam, Roopika, and Alex Gil. 2022. “Introduction: The Questions of Minimal Computing.” Edited by Alex Gil and Roopika Risam. *Digital Humanities Quarterly* 16 (June). <http://digitalhumanities.org/dhq/vol/16/2/000646/000646.html>.
- Rockwell, Geoffrey. 2006. “TAPoR: Building a Portal for Text Analysis.” In *Mind Technologies: Humanities Computing and the Canadian Academic Community*, edited by Raymond George Siemens and David Moorman, 285–89. Calgary: University of Calgary Press. <https://www.deslibris.ca/ID/415538>.
- “Starting Fresh: The Wikibase REST API.” 2023. Wikimedia Tech News. September 7, 2023. <https://tech-news.wikimedia.de/2023/09/07/starting-fresh-the-wikibase-rest-api/>.
- VanderWal, Thomas. 2004. “Feed On This.” October 3, 2004. <https://www.vanderwal.net/random/entrysel.php?blog=1562>.
- Vismann, Cornelia. 2009. “Arché, Archiv, Gesetzesherrschaft.” In *Archivologie: Theorien des Archivs in Wissenschaft, Medien und Künsten*, edited by Markus Knut Ebeling, Stephan Günzel, and Aleida Assmann, 89–103. Berlin: Kulturverlag Kadmos.
- “Wikidata:WikiProject DH Tool Registry.” 2024. Wikidata. November 24, 2024. [https://www.wikidata.org/w/index.php?title=Wikidata:WikiProject\\_DH\\_Tool\\_Registry&oldid=2277445090](https://www.wikidata.org/w/index.php?title=Wikidata:WikiProject_DH_Tool_Registry&oldid=2277445090).
- Zhao, Fudie. 2022. “A Systematic Review of Wikidata in Digital Humanities Projects.” *Digital Scholarship in the Humanities*, December, 1–22. <https://doi.org/10.1093/lc/fqac083>.