Datenbank-Spektrum (2025) 25:167–178
https://doi.org/10.1007/s13222-025-00515-7

SCHWERPUNKTBEITRAG

A Flexible Search System for Integrated Authority Data—ADISS

Leon Fruth[1] · **Tobias Gradl**[1] · **Maximilian Hebeis**[1] · **Andreas Henrich**[1]

Received: 14 June 2025 / Accepted: 6 October 2025 / Published online: 6 November 2025

Abstract
This paper presents ADISS, a generic and flexible search system designed to integrate heterogeneous authority file providers. Authority data is used to unambiguously identify entities such as persons, places, and institutions. As individual data providers typically do not offer both broad, universal coverage and domain-specific, in-depth data, many real-world application scenarios require combined access to multiple datasets. In the context of Digital Humanities, combining multiple authority data providers helps to improve the coverage and to resolve ambiguities more effectively during data curation. Instead of relying on separate endpoints to access multiple data sources, we address this need by aggregating the datasets and exposing them through integrated query and result models, thus simplifying access at the level of technical interfaces and schema integration. We present a highly configurable search API, which offers a diverse range of search and filtering options, to leverage the integrated data sources. We demonstrate its practical value through two active projects, that require semi-automatic retrieval, as well as user-centered search scenarios for different authority file providers. An evaluation based on manually curated data shows high retrieval accuracy, highlighting the system's effectiveness. We further illustrate how ADISS's generic and highly configurable design makes it adaptable and reusable for a wide range of use cases, and we outline directions for future improvements.

Keywords FAIR · Digital Humanities · Authority Data · Data Curation

1 Introduction

The FAIR acronym, which stands for Findability, Accessibility, Interoperability, and Reusability of digital assets, has become an important concept in academia, serving as a crucial criterion for effective research data management [1]. Authority files play a key role regarding the implementation of the FAIR principles by enabling unambiguous identification of entities like locations, institutions or persons. Rather than relying solely on textual descriptions, links to entity representations in knowledge bases enrich and contextualise research data and thus enhance their overall findability and interoperability, making it easier to discover and use [2].

There are many sources for authority data, each containing distinct sets of entities and focusing on different entity attributions that are influenced by the legal and contextual framework of the respective provider. These providers can be broadly categorised into three groups. The first group consists of *highly regulated authority file providers* such as the Gemeinsame Normdatei (GND) or the Virtual International Authority File (VIAF) that are managed by trusted institutions and libraries. The GND is a project of the German National Library (DNB) and is strictly maintained within quality standards and legal settings [3]. The VIAF is a collaborative effort by the libraries DNB, the Library of Congress (LOC) and the Online Computer Library Center (OCLC) to create an overarching authority file which spans multiple national libraries and archives [4]. Due to their focus on quality and institutionally induced biases[1], providers like the GND expose sets of highly regulated and curated entries that are naturally limited in quantity.

✉ Leon Fruth
leon.fruth@uni-bamberg.de

Tobias Gradl
tobias.gradl@uni-bamberg.de

Maximilian Hebeis
maximilian.hebeis@uni-bamberg.de

Andreas Henrich
andreas.henrich@uni-bamberg.de

[1] Media Informatics Group, University of Bamberg, An der Weberei 5, 96047 Bamberg, Germany

[1] National and cultural biases in the case of national libraries due to their mission.

The second group includes *community-based projects* such as Wikidata which hold substantial amounts of user-generated data. Wikidata is commonly utilised in Digital Humanities (DH) projects to support metadata curation, annotation and Named Entity Recognition (NER) [5]. Additionally, Wikidata is also used in natural sciences to improve data integration and accessibility by identifying people that are associated with entities, such as specimens and taxonomic names [6, 7]. Despite unmatched degrees of quantity, these community-based platforms pose challenges related to data accuracy and exhibit biases driven by public interests [5].

Lastly, the third group consists of highly-focused, *specialised data repositories*. For instance, Memorial Archives contains entity descriptions in the context of the Flossenbürg concentration camp, while the historic place names in Bavaria from Geschichte Bayerns can be a valuable resource for certain research communities. Often being very detailed in their respective contextual setting, specialised data repositories attempt to fill specific gaps, but do not contain descriptions of entities that are not relevant within their specific scope.

As a result of these limitations, relying on a single authority data provider often does not satisfy the requirements of DH projects. Integrated access to multiple data providers on the other hand improves possibilities for referencing and curation—enhancing the quality of the structure, detail, and semantic richness of the data and thus its findability [8].

In this paper we introduce the Authority Data Integration Search System (ADISS), which integrates various heterogeneous authority file providers and provides a wide range of search and retrieval techniques. Its search API is based on preliminary work in the infrastructural frameworks of DARIAH-DE and CLARIAH-DE [9]. The API's highly configurable nature allows to dynamically accommodate different data schemas and diverse retrieval requirements. For two use cases (the NFDI consortium Text+ and the project Oral-History.Digital), we present an optimised search query configuration tailored to the unique requirements. One of these configurations is used for an evaluation, giving insights into existing challenges and potential improvements.

The remainder of the paper is organised as follows: The requirements on integrated access to several authority file providers in the DH will be elaborated further in Sect. 2 based on our two use cases. Section 3 provides an overview and classification of the related state-of-the-art. Section 4 presents the selection and characteristics of the initial data providers currently integrated into ADISS. Section 5 discusses architectural considerations and the chosen unified indexing approach for efficient data integration and retrieval. The implementation details, including data preprocessing, mapping into a homogeneous schema, and the configurable search API, are described in Sect. 6. Section 7 illustrates the application of ADISS through two active use cases in the DH domain. A first evaluation of the system's retrieval accuracy using real-world data from one of our use cases is presented in Sect. 8. Finally, Sect. 9 concludes the paper with a summary of the contributions and outlines directions for future improvement.

As an expansion of our previous workshop publication [10], the following additional contributions are given in this paper: A more comprehensive consideration of related work. The inclusion of a use case section detailing the current state of the two active projects—Text+ Registry and Oral-History.Digital (oh.d)—that demonstrate ADISS's applicability in real-world DH workflows. An evaluation section presenting quantitative results based on 883 queries and providing insights into the system's precision, challenges, and areas for improvement. Additional details, notes, examples, and explanations throughout the paper.

2 Motivation

The idea of consolidating access to multiple authority file providers results from the informational needs in the research infrastructures Text+ and oh.d. Despite differences in contexts, user communities, and usage scenarios, both initiatives share the goal of enhancing the visibility and findability of research data, and have similar requirements related to the retrieval and use of authority data.

The project oh.d develops and operates a data curation and research platform for collections of audio-visually recorded narrative interviews. The metadata of these interviews typically include attributions in relation to entities, such as the names of interviewer and interviewee and other referenced persons, institutions and locations, often historical places where the interview was conducted or to which the interviewee is connected. Transcriptions are available for a majority of the interviews in oh.d and include large amounts of automatically and manually extracted named entities that should be linked to authority data. Interviews in the context of oh.d are mostly life history interviews, which—while often focusing on the individual's entire life—are linked to a specific thematic context, such as the Holocaust in Germany or life as a migrant worker. Access to multiple data providers facilitates referencing of both prominent and very specific entities which are typically not found within the dataset of one single provider. Furthermore, some of these data providers do not offer features such as consistent coordinates or labels in multiple languages, which are functionally required in the oh.d portal. It is typical for oh.d that interview data is imported from existing databases and manually edited in the oh.d portal,

which requires support for authority data search scenarios in both semi-automated and manual settings.

Text+ is a consortium in the German National Research Data Infrastructure (NFDI). It focuses on language and text resources that are of high relevance in related disciplines, and aligns them along the categories of digital collections, lexical resources, and editions. A central component of the Text+ infrastructure is the Text+ Registry, which serves as a central resource catalog. It builds on resource descriptions scattered across existing catalogs and data sources and facilitates the manual addition and enrichment of metadata. A fundamental aspect of enrichment is the contextualisation of resources, i.e. the explication of references between resources (e.g. the edition of a particular letter being part of the complete edition of all works of an author), and to related entities such as persons, institutions and locations. Much like the oh.d portal, the Text+ Registry requires support for semi-automatic correlation of entities when importing metadata from existing catalogs. First the entities are imported automatically using the metadata, which can be names of entities or identifiers from an authority file, then the resolved entities—especially when resolving by name—need to be manually confirmed for correctness. Furthermore, imported entries and resources missing in connected catalogs are manually curated by domain experts and require user-centered search facilities for suitable authority data.

Both Text+ and oh.d implement functionalities to contextualise data by means of identifying and explicating relations to various classes of entities. Focusing academic contexts and users, both infrastructures prefer information of high-quality and domain-specific authority file providers, but require their combination with community-based sources to be able to search in a large set of entities. As this combination requires processing of multiple query and response formats and the identification of duplicates between sources, an implementation in individual project contexts seems redundant and impractical. With requirements for both semi-automatic and user-centered authority data search scenarios, the presented use cases are examples for projects with similar requirements, which allows ADISS to be designed to support the specific needs of these projects and to be generic and reusable.

3 Related Work

Authority control systems were first implemented by single institutions in order to provide controlled vocabularies to their resource catalogues. As the authority files grew in regional and thematic scope and evolved to general providers of entity and identity management, they are now seen as providers of linked data [11]. Thus, authority records often contain references to corresponding entities from other authority data providers via predicates such as `owl:sameAs` [12].

Multiple prior efforts rely on authority data from multiple providers [5]. As previously stated, authority data providers often differ in their conceptual and informational focus, making the integration of multiple authority files necessary for use cases which span multiple domains or aim to achieve more comprehensive annotations. For example, Adams [13] draws on multiple source authority files to facilitate searching for historical events by combining historical and geographical authority data, while Koch et al. [14] integrate records from Wikidata and DBpedia records to enrich existing archival data.

Although the use of multiple authority files has shown clear benefits, research into integrating access to authority data across different databases remains limited and often tailored to specific use cases. Existing studies on authority file reconciliation primarily focus on aligning smaller databases with large-scale authority files [15, 16]. Ravelli and Mataloni [17] describe an integrated search system for the Italian National Library Service (SBN). However, this was achieved through prior manual reconciliation of several smaller authority files, rather than cross-database integration.

To date, the aforementioned VIAF represents the most comprehensive and widely adopted effort for interlinking multiple large-scale authority files, with other initiatives like the International Standard Name Identifier (ISNI) being largely built on top of VIAF [18]. Despite of this importance, the inner workings of the VIAF matching and clustering of authority records from different sources remain proprietary aside from a few articles describing narrow aspects of the matching process [19, 20]. Additionally, the VIAF follows a top-down approach, incorporating only authority data from trusted highly regulated authority file providers, with Wikidata being the sole exception [21]. Other community-based projects like GeoNames or OpenStreetMap (OSM) are excluded from VIAF.

4 Data Providers

The selection of the initial set of authority file providers for the implementation of ADISS has been influenced by the needs of the usage scenarios described in Sect. 2. To determine the general feasibility, scalability and robustness of the proposed solution, we have initially focused on the distinct entity type of locations and gathered data from multiple providers—commencing with the relevant and large datasets of GND, Wikidata, OSM and Geonames.

Due to its national and academic setting, the central authority file provider used in the previously described use

cases is the GND provided by the DNB. The GND consists of six different entity types and contains over 10 million total entities.[2] The data can be obtained as a data dump or queried via a web-API. The API allows entities to be retrieved by their identifier and through a search function, which includes basic filtering and sorting options. However, it lacks features such as fuzziness, suggestions and geographical filters. Additionally, the data does neither fully support multilingual names and descriptions nor does it contain sufficient geographical information for our requirements. However, other authority data sources can be used to fill these gaps for entities that exist in multiple sources and contain at least a unidirectional reference to the respective GND entry.

A rather large data source is Wikidata with over 100 million total entities. Within its extensive collection, Wikidata offers a high quantity of labels, aliases, and descriptions in numerous languages along with annotated language codes. This comprehensive dataset provides access to a wealth of information that cannot be found on a comparable scale in other sources such as the GND or similar providers. Moreover, Wikidata also features a large amount of reference links and identifiers connecting it with various other websites and authority data providers, including VIAF, GND, OSM and Geonames. These connections enable the linking of entities from different sources which enhances overall knowledge integration. It should to be noted that, due to their community-based nature, Wikidata and other providers have some quality issues, such as (near) duplicate entries or incorrect links to other authority data providers [22]. On the other hand, widespread use in the DH shows that these problems are clearly offset by the strengths of Wikidata for many use cases.

The data from Wikidata can be accessed through dumps in multiple formats as well as via a SPARQL-endpoint or the Wikibase REST-API. Like the GND API, these endpoints do not support options like fuzziness or geographical filtering at present.

An additional data provider utilised for the described use cases is OSM, which includes detailed coordinates of locations and regions but overall with varying quality of data. OSM data can be queried through different APIs. A prominent example is Nominatim, which offers text searches, but again lacks functionalities such as fuzziness and geographical filtering. Another option can be found in Photon, a service that allows some geographical prioritising, but does not include many language labels that are present in Nominatim.

GeoNames is another geographical database that contains over 11 million places. The data is accessible through a web API that offers full-text search capabilities; however, it misses certain features like the aforementioned APIs. Both OSM and GeoNames offer complete datasets as downloadable dumps for further usage.

[2] As of 27-05-2025.

5 Background and Architectural Considerations

Building a system that allows querying the above mentioned datasets in an integrated manner can be approached in different ways. One option is implementing a federated meta-search that queries and aggregates over multiple of these endpoints. However, this approach is limited by the capabilities of the underlying APIs, lacking in required search functionalities necessary for the previously described use cases. Moreover, it is very difficult to derive a comprehensive, consolidated ranking based on search results from different APIs due to differences in data types, query languages and retrieval models between providers.

The second approach, which is utilised in this work, involves gathering the data from multiple data sources and integrating it into a homogeneous data model. By indexing the data, we can offer extensive search functionality through a single API while accessing data of various sources concurrently. This method provides arguably better performance than the previously mentioned approach as it reduces network latency by minimising requests down to only one instead of $N + 1$ requests with N queried data providers.

Moreover, heterogeneous data is integrated into a homogeneous format and cached during indexing rather than during runtime when the providers are queried, which further improves the performance. Additionally, highly specialised data sources that lack an dedicated API can be added, offering research projects requiring such information an endpoint to query the data.

For scalability and reduced maintenance efforts, a hybrid approach combining download and API-access could also be used, which has been presented in [23]. Here an upstream search API has been created for some providers while existing endpoints have been used for others. Similarly, the GFBio Terminology Service integrates various terminologies either internally indexed or externally accessed [24]. In general, for some providers without a sufficiently powerful search interface, a new central index can be created on the basis of the downloads provided, combined with a meta-search for other providers with large data sets and sufficiently powerful search functionality. Such an approach could also be considered for ADISS in the future if scalability and up-to-dateness pose a problem, or if a data provider does not offer its data for download but allows API-based search access.

Currently, the data for ADISS is regularly updated to meet the needs of the above usage scenarios, and the com-

bined storage requirement for the four integrated databases is approximately 67 GB, with individual sizes as follows: GND—6.1 GB, Wikidata—29.8 GB, OSM—27.8 GB, and GeoNames—2.93 GB.

Based on these system design considerations the following section outlines the implementation for our proposed solution ADISS, not only addressing the requirements from oh.d and Text+, but also presenting a generic solution expandable to other fields and use cases.

6 Implementation

This section describes central aspects of the implementation of ADISS. Figure 1 displays the architecture of the service. For each data provider, a dedicated data wrapper handles the corresponding format and structure. The initial three stages of the data pipeline are discussed in subsection Sect. 6.1. Subsequently, we outline the concept for the integrated data model in subsection Sect. 6.2. Finally, in subsection Sect. 6.3 the Generic Search (GS) API is presented with its most relevant features and configuration options, providing search and retrieval functionalities for the processed data.

6.1 Data Collection and Processing

Each data provider considered within the application context offers download options for its authority data. For every provider, a dedicated data wrapper is implemented in Java, utilising the Spring Boot Framework. While these wrappers share a common set of core functionalities, they are tailored to the respective structural constraints of each provider's data. Furthermore, existing tools and libraries, such as osm2pgsql for processing OSM data dumps, and the Wikidata Toolkit for handling Wikidata, have been integrated to streamline the data processing workflows.

The process begins by downloading data dumps from their respective source in the available formats: GND entities are retrieved in JSON-LD, OSM data in PBF, and Wikidata in JSON or XML. Following the extraction, the data is filtered to retain only entries relevant to our intended use cases. For instance, in the case of wikidata, entities that are subclasses of irrelevant categories, such as astronomical objects or chemical compounds, are excluded. Similarly, unnecessary attributions and media content are removed. For OSM, entries such as highways or power lines are disregarded. This preprocessing step substantially reduces the volume of data to be processed, particularly for Wikidata and OSM.

Moreover, certain fields require structural normalisation. For example, coordinate data, originally represented in various formats such as Well-Known Text (WKT), are uni-

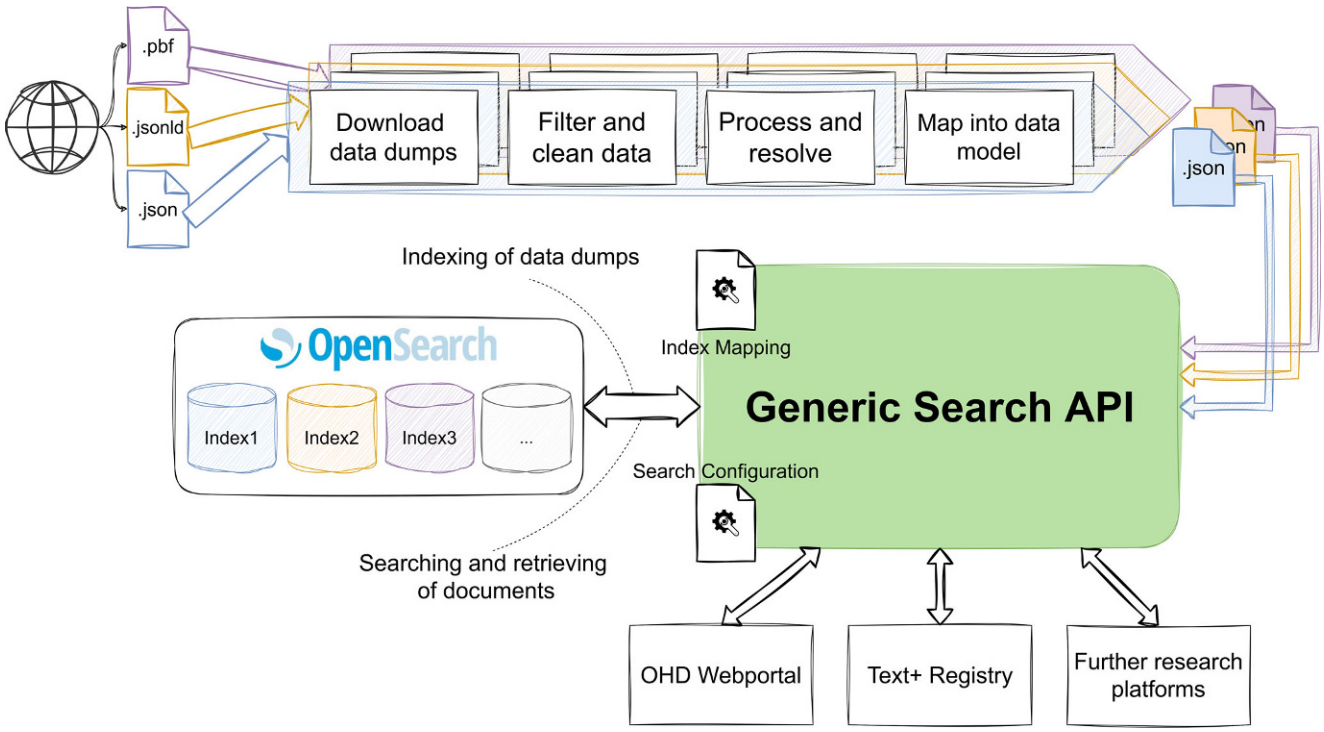

Fig. 1 Conceptual architecture of the Authority Data Integration Search System (ADISS)

formly converted into the GeoJSON standard. To facilitate comprehensive search capabilities on data from multiple sources, similar harmonisation needs to be applied to other data fields like the entities *type* or *classes* of individual data providers and country- or subdivision-codes, which are present in different formats. For example the GND offers references of their geographic area codes to the MARC list for countries, which will be utilised in the future.

6.2 Data Model

After data downloading and preprocessing, a subsequent task for data wrappers consists in the alignment of source data with our internal data model, which is designed to contain data in its original *and* integrated representations. Hence, the internal data model is composed of two main parts:

1. `resource`: This dynamic field holds the original source-specific data.
2. `integration`: Used for efficient full-text searches and filtering by abstracting and transforming the original data into an integrated format. An example is provided in Fig. 2.

```
{
  "_integration": {
    "names": {
      "name": {
        "@value": "Bamberg",
        "@language": "en"
      },
      "variantNames": [{
        "@value": "Bamberg",
        "@language": "de"
      }]
    },
    "provider": "wikidata",
    "coordinates": { // ...
    },
    "id": "Q3936",
    "type": ["college town", // ...
    ],
    "descriptions": [{
      "@value": "kreisfreie Stadt...",
      "@language": "de"
    }],
    "url": "http://www.wikidata.org/e..."
  }
}
```

Fig. 2 Simplified example for the integration field of an entity

With the `integration` field being syntactically and structurally consolidated, consuming services can formulate faceted queries over multiple data sources.

As a final result of data preprocessing, data is saved in an intermediary JSON format. This step allows us to functionally decouple wrappers from the indexing logic. New wrappers for other data sources only need to output data conforming to the constraints of our JSON format, which are then indexed through our API.

6.3 Search and Retrieval

The Genereic Search (GS) API, utilised for indexing and retrieving the data, utilises OpenSearch as backend. Individual entities can be accessed through a GET-request by specifying the index name of the data source and the entity's identifier. A key feature of the GS API is the flexible search function that supports not only the diverse requirements described in Sect. 2, but can also be dynamically adapted towards new use cases.

The GS is highly configurable via configuration files, which allows to enable, disable and parameterise different OpenSearch functionalities and to consolidate them in terms of *search profiles* that can be tailored to the specific needs of use cases. A search query may be constructed by combining multiple query types, for example full-text queries that employ Lucene's Query Parser Syntax, as well as queries for (near-)exact matches in keyword fields. To accommodate the diverse search scenarios described in Sect. 2, different combinations of query types are applied to the relevant fields. The semi-automatic search scenario when bulk-importing catalog data has a focus on high-precision and uses exact match queries on entity identifiers and URLs among others. The user-centered manual search focuses a ranking mostly based on full-text queries for high-recall results. Notably, additional configurations, such as aggregations, search suggestions, keyword highlighting, and the specification of geographical fields for utilising distance or polygon filters can be specified in the search profiles. In a user-centered search scenario, for instance, keywords may be aggregated to provide the user an overview of the results and act as facets, while query suggestions are also displayed.

Another important feature within the search configuration is the dynamic alteration of the result relevance score by using custom scripts. This is currently employed to favour entities that contain links to matching entities of other providers, which we interpret as an indicator of the general relevance. These entities are preferred over others with a similar score but less links. Although experimental, early results indicate the applicability of this approach, and we will methodologically invest into this feature in the future to mitigate risks of e.g. hiding specialised entries.

The API can be configured to use different relations between entities. Presently, only `sameAs` relations between matching entities from different data providers are considered. Retrieving these related entries during the search is accomplished with one additional request on the basis of entity identifiers, resulting in a minimal performance impact. The future roadmap includes enabling the specification on how these matching entites across multiple sources can be merged and which fields from which sources to prefer when conflicts occur. Currently, the search results are returned as a list of the matching entities instead of merging them.

The options described above for parameterising a search request can be specified in terms of a `POST` request body, showcased in Fig. 3. The request considers all indices that contain geographical data (line 3). It looks for entities that contain coordinates (line 5), contain the country code "DE" (line 6) and are within 10 km of some coordinate (line 8–9). The linked entities of the given results are joined in the result list (line 10). Another request is shown in Fig. 4, using "Wolfgang Goethe" as query string to search through person entities of the GND.

```
{
  "query": "Bamberg",
  "indices": ["geo"],
  "filter": {
    "exists": "_integration.
      coordinates",
    "_integration.countryCode": "DE"
  },
  "coordinates": [49.9031, 10.8695],
  "distance": 10000,
  "joinLinks": true
}
```

Fig. 3 JSON body for a `POST` search request to https://c102-142.cloud.gwdg.de/adiss-gs/search/default using the search profile "default"

7 Use Cases

The projects outlined in Sect. 2 remain under active development, and the requirements that motivated ADISS's design have not yet been fully realised. Meanwhile, the volume of resources and interconnections within both project initiatives grows steadily. This section therefore provides an overview of ADISS's current status in the context of these two research projects.

7.1 Oral History Digital

The project oh.d focuses on importing geographical authority data but is planned to be expanded to additional entity types, such as persons. Currently all entities are manually imported by the curators in their portal. The portal hosts 4137 interviews as of 13-06-2025. Multiple data providers are considered for oh.d: the GND, OSM, Wikidata and GeoNames. Each provider contributes unique value, for example, the GND includes historically significant locations, while OSM offers highly granular data such as street names and good coordinate coverage. To address multilingual and geospatial requirements, the system leverages external links between corresponding entities across these sources for enrichment. The aggregated data is mapped and transformed using the DARIAH-DE Data Modeling Architecture (DME) and the Transformation Service (TS) introduced in [23], enabling integration into the oh.d infrastructure.

The search profile created for this use case allows coordinate filters, such as range and polygon queries to enable curators to intuitively filter the results using a geographical map. Search queries take into account all available names, including multilingual, alternative, and historical variants. Additionally, data can also be retrieved by supplying an identifier or the URL of an entity. To further improve the search results, it has indicated effective to boost the scores of entities, using the number of external links to other authority file providers. This improved the accuracy of the search results in both of our use cases.

7.2 Text+ Registry

Entities used for contextualisation in the Text+ Registry are currently limited to institutions and persons. Future plans involve the incorporation of more entity types, such as works, keywords and time periods. The only provider presently utilised is the GND. Since not all required entities are available this will be expanded in the future, by incorporating further providers, such as Wikidata and ORCID to increase the quantity of available entities. This is especially the case with persons, which are related to a collection or edition within Text+, but unknown to the GND.

As of 01-06-2025, there are 839 person entities and 148 institution entities imported and linked to resources in the Text+ Registry. Figure 5 illustrates the connections between

```
{
  "query": "Wolfgang Goethe",
  "indices": ["gnd_person"]
}
```

Fig. 4 JSON body for a POST search request to https://c102-142.cloud.gwdg.de/adiss-gs/search/registry using the search profile "registry"

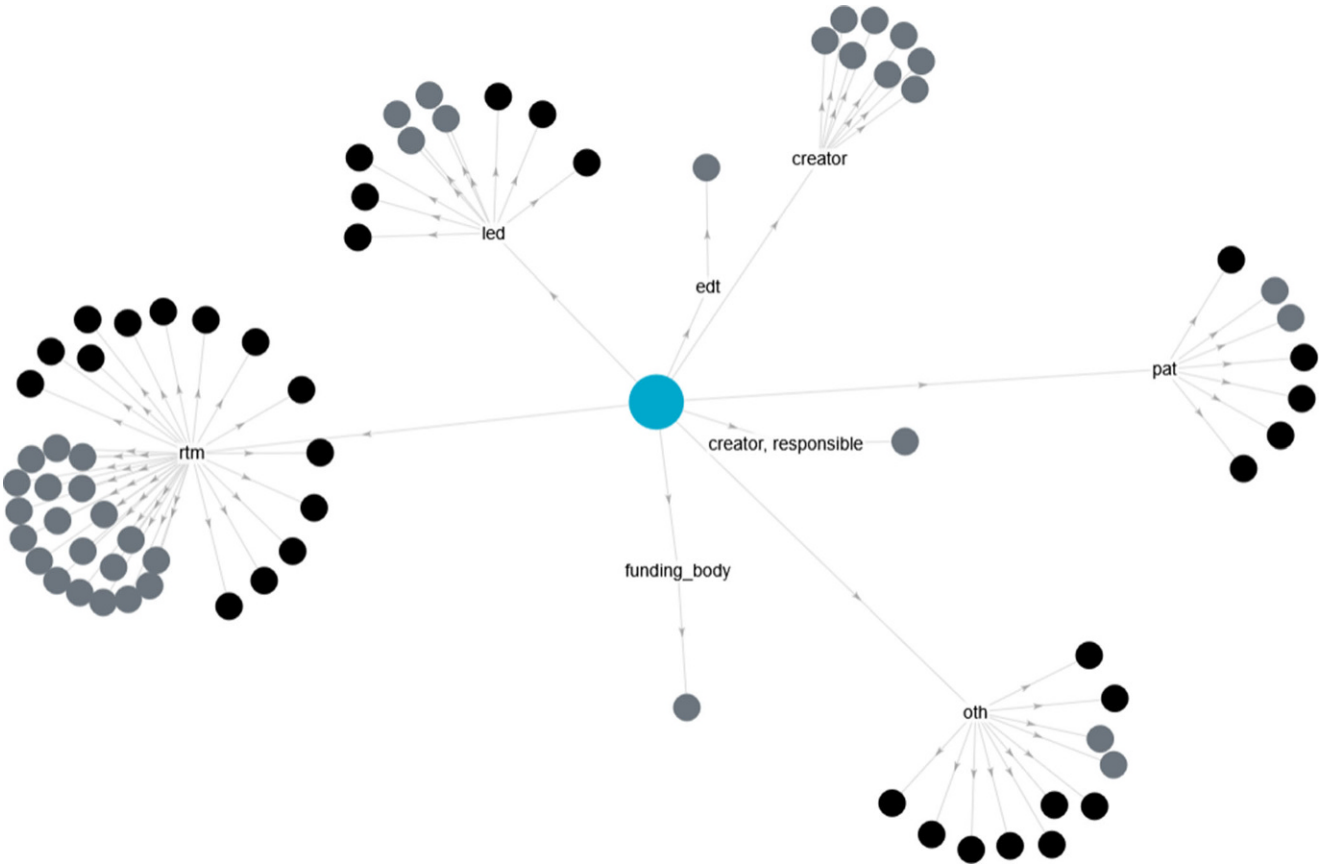

Fig. 5 Graph visualisation of an entry in the Text+ Registry Frontend. The blue circle represents an edition resource that is connected to grey and black circles, which denote person and institution entities. The labels on the connections (e.g. editor [edt], research team member [rtm], patron [pat], lead [led], other [oth]) are codes from the MARC code list for relators: https://www.loc.gov/marc/relators/relaterm.html

a sample edition and its associated person and institution entities. In this visualisation, grey circles represent entities successfully resolved via ADISS, while black circles denote person entities that have not yet been resolved.

Entities are either imported automatically or manually. Resources automatically imported into the Text+ Registry often contain references to entities within an authority file or names of entities which are used to query ADISS automatically. When the score of the first result is considered sufficiently high in itself and twice as high as the second result, the first result is automatically linked to the respective resource. If the confidence level is not reached, the entity requires manual disambiguation. However, this process can yield errors, for example, when a searched person is not available within an authority file, but another individual with the same name is available and reaches the confidence level. Manual reviews are therefore still required, to ensure accurate metadata.

When manually entering an entity, the user queries the API using a text field, where the top 10 results are returned (see Fig. 6). The search profile is similar to the previously described profile for oh.d. It does not consider any geographical filters, but also searches through all available names and considers identifier and URL matching. Curators in this context also search using abbreviated names, such as "IDS" to search instead of "Institut für Deutsche Sprache", so abbreviated names are also considered in the Registry's search profile. The manual search capabilities have been evaluated in the following chapter.

8 Evaluation

To evaluate the performance of ADISS, we constructed an evaluation dataset using manually curated entries with associated authority records from the Text+ Registry. Although the Text+ Registry supports both automatic and manual imports, this evaluation focuses exclusively on manually curated entries, where curators explicitly selected person or institution entities from the GND for enrichment. The goal with this evaluation is to show the retrieval effectiveness for curators using ADISS' search API on GND data.

In this section, we begin by further describing the evaluation dataset. We then present and discuss the ranking re-

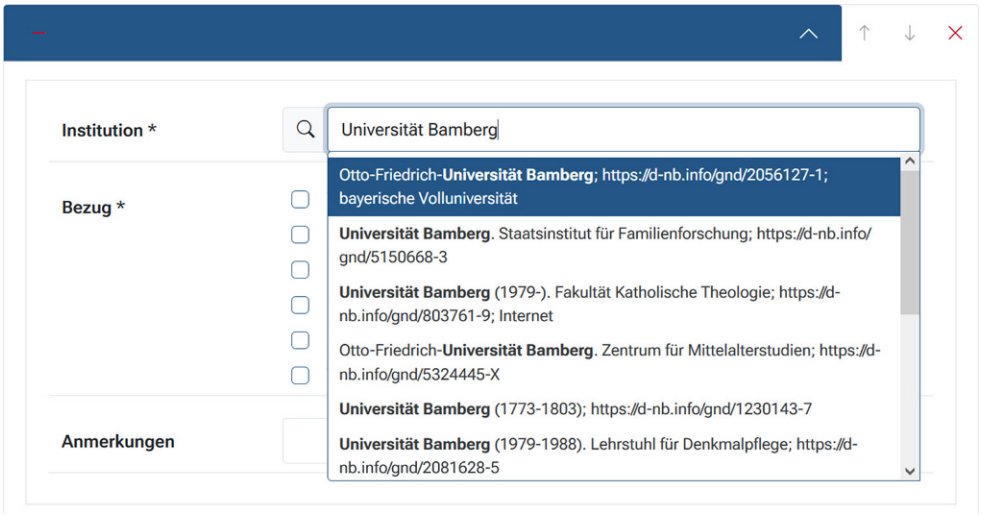

Fig. 6 Querying ADISS in the Text+ Registry curation UI. The user can select one of the top 10 search results and then choose the type of relationship (*Bezug*) based on MARC relator codes to the edition under consideration. Explanatory comments (*Anmerkungen*) on the relationship are also possible

sults of ADISS, highlighting specific challenges observed in cases where the correct entity was not successfully retrieved. In addition, we analyse a subset of entries from the Text+ Registry that lack a GND reference, although ADISS was available as a supporting tool for connecting to the GND when manually curating the data, to gain a deeper understanding about potential reasons for missing references even with tool support. For clarity, it should be noted that these entries were not included in the evaluation dataset, due to their missing references.

8.1 Dataset

Since persons and institutions are the only entity types currently handled by the Text+ Registry, our evaluation dataset is restricted to these. To construct the dataset, we considered all entities manually added to edition resources. For each of the entities, the respective GND identifier was treated as the ground truth and the entity's names were used as the query inputs. Person entities were queried using only their primary name in the experiments presented in subsection Sect. 8.2, whereas for institution entities additional queries were generated using all available variants and abbreviated names. Because the entity names and identifiers were manually curated, the evaluation dataset may of course contain human errors—but overall it corresponds to a high current scientific standard.

Table 1 gives an overview over the evaluation dataset, with a total of 267 entities and the resulting 953 queries. Each query will be used for evaluation in the following subsection.

The results on this test case offer a preliminary insight into the precision of ADISS. However, in real-world searches, users often supply approximate or partial names rather than exact canonical labels. A future study will leverage actual user-supplied queries to evaluate and refine our search profiles.

8.2 Ranking Evaluation

To assess the ranking precision, we emulated the search setting used for the Text+ Registry, utilising its search profile and returning the top 10 search results. In this use-case the retrieval of a single correct target entity is required. Therefore, we report the Mean Reciprocal Rank (MRR), together with the percentages of queries with the correct GND record as the first result (*Top 1*) and within the *Top 10* results. These measurements reflect the two goals resulting from the UI shown in Fig. 3: Ideally, the correct record should be the *Top 1* result, but at least it should be within the *Top 10* results shown in the result list.

Table 2 displays the evaluation result, where 98.32%% of all entities could be retrieved in the top 10 results. These results indicate strong overall ranking performance across all queries in this evaluation setting.

In total, sixteen queries (five for persons, eleven for institutions) failed to return the correct entity in the top 10 results (1.68% of the 953 queries). All five person-query failures involved name ambiguities in the GND and lacked sufficient disambiguating context. For instance the com-

Table 1 Overview of queries in the evaluation dataset for persons (P) and institutions (I). For the person entities only the primary names (PN) are used as queries, whereas with institution entities the variant names (VN) and abbreviated names (AN) have been used in addition.

	P	I	Total
Entities	194	73	267
PN	194	73	267
VN	–	640	640
AN	–	46	46
Total Queries	194	759	953

Table 2 Evaluation results: Mean Reciprocal Rank (MRR), and percentage of queries where the ground-truth entity is ranked first (Top 1) or within the top 10 (Top 10). Results are reported separately for person and institution entities, considering the primary names (PN), variant names (VN) and abbreviated names (AN).

{}	Queries	MRR	Top 1 (%)	Top 10 (%)
Person PN	194	0.9363	91.24	97.42
Institution PN	73	0.9121	84.93	100.00
Institution VN	640	0.8682	80.00	98.28
Institution AN	46	0.7451	63.04	100.00
Total	953	0.8795	81.84	98.32

mon German name "Thomas Müller" returns several hundred individual person entities in the GND that share the same name.

All eleven failed institution queries were variant name queries for five individual entities. For instance, querying "University of Freiburg" instead of its German primary name "Albert-Ludwigs-Universität Freiburg" led to matches with departmental sub-entities (e.g., "University of Freiburg. Genetics & Experimental Bioinformatics"), which had the English name as primary name. This behaviour reflects the current ranking bias toward primary names and exact textual matches.

8.3 Analysis of Missing GND References

Besides the evaluation dataset with GND references, we considered the 45 curated records in the Text+ Registry that lacked a GND reference and could therefore not be used in the gold standard. These records contained only the user-supplied name as a textual label, although ADISS was available to connect to the GND—a result that may stem from various factors. We can classify the possible reasons into four categories:

1. **Data absence error**: The entity is not represented in the authority dataset.
2. **Stale data error**: The entity exists in the authority dataset, but the data within ADISS is outdated.
3. **Retrieval error**: The entity is present within ADISS, but failed to be retrieved successfully.
4. **Human error**: A human error during data curation led to incorrect or missing information.

Out of the 45 entities, there was only one institution entity, the "Niedersächsisches Ministerium für Wissenschaft und Kultur". However, it is available in the GND and is correctly retrieved as the top result in ADISS when querying its primary name. This case might reflect either of the previously described errors, for example if the entry was not available within the GND at the time of curation, it could be a data absence error.

To better understand the 44 missing person references, we manually examined a random subset of ten entities using both ADISS and the GND Explorer to query the missing persons by their name. In six cases, the queried names could not be found in either system. In the remaining four cases, the name appeared as the top result in ADISS. However, due to name ambiguity, the correct identity of the requested entity could not be reliably confirmed without additional insights of the curator(s). These examples show challenges, regarding entity availability and disambiguation. The availability in the GND cannot be influenced by ADISS, while more powerful approaches to disambiguation are an important field for future work.

The evaluation results demonstrate that ADISS performs strongly in retrieving relevant authority entities, with nearly 99% of all queries successfully returning the correct entity within the top 10 results in subsection Sect. 8.2. Our analysis highlights challenges related to entity availability, outdated data, and name ambiguities. These findings show the system's current effectiveness in supporting data curation workflows while highlighting the importance of improving name disambiguation and expanding the data coverage, especially for person entities.

Despite the strong results, this evaluation has clear limitations. It is restricted to the GND as the sole authority file and to curated queries from the Text+ Registry, which cover only a limited range of entity types. Moreover, the study is entirely system-focused only and does not include real-world user query inputs, thereby providing limited insights into its practical usefulness in curation workflows. Future evaluations should therefore broaden the scope to include multiple authority providers, incorporate real-world user queries, and gather curator feedback to assess ADISS' performance more comprehensively.

9 Conclusion and Future Work

This work introduced ADISS, a search system designed to integrate multiple authority file providers and support the retrieval of enriched authority data for diverse data curation workflows. Motivated by the requirements of two ongoing research projects in the field of DH, that currently utilise the API, we have demonstrated how our generic and configurable approach enables flexible methods querying and retrieving data from heterogenous authority data sources.

The system is under active development, with several enhancements planned, including the integration of additional data providers. For example, Memorial Archives and historic place names from Geschichte Bayerns will expand the coverage of historical and specialised location and person entities, that are not available in more general-purpose datasets. Person records from ORCID will extend the set

of person entities of the GND and Wikidata, though with a notable research-focused bias due to ORCID's academic focus. Further incorporating temporal authority data, for example from PeriodO, will enhance the utility of ADISS in historical research contexts [25]. To improve retrieval performance, we plan to implement language-specific text processing, including custom stopword lists and stemming algorithms to accurately derive word stems in the respective language. Additional techniques will include the decomposition of German compound words (e.g., "Sprachwissenschaft" is separated into "Sprache" and "Wissenschaft") or removing elisions for French texts (e.g., "l'avion" is processed to "avion").

We are also exploring the adoption of interoperability standards, such as the Schema.org schemas and the DataCite metadata schema, to ensure better interoperability [26].

The introduction and fine-tuning of various search profiles tailored to specific use cases have demonstrated promising improvements in search effectiveness and flexibility. An initial evaluation based on manually curated entities from the Text+ Registry has yielded promising results, indicating that ADISS is effective in retrieving relevant authority records. Nonetheless, future work will involve more comprehensive and systematic evaluations, especially incorporating real user queries and interaction logs. These studies will be conducted use-case-specific, within both the Text+ and oh.d project contexts.

Funding This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). The authors appreciate the financial support. The work was done in the context of the NFDI consortium Text+ and the project Oral-History.Digital 2. Project number Text+: 460033370; Project number Oral-History.Digital 2: 437972564.

Funding Open Access funding enabled and organized by Projekt DEAL.

Supplementary information The source code and configuration files related to ADISS are publicly available:

Authority Data Wrappers: https://gitlab.com/minfba/resinfra/adiss

GS API: https://gitlab.com/minfba/resinfra/generic-search/generic-search-api

ADISS Configuration: https://gitlab.com/minfba/resinfra/adiss/adiss-config This repository contains search profiles, schema mappings, and additional configuration for the GS API.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, Santos SLB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Lei J, Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3(1):160018. https://doi.org/10.1038/sdata.2016.18
2. Hawkins A (2021) Archives, linked data and the digital humanities: Increasing access to digitised and born-digital archives via the semantic web. Arch Sci 22:319–344. https://doi.org/10.1007/s10502-021-09381-0
3. Behrens-Neumann R, Pfeifer B (2011) Die Gemeinsame Normdatei: Ein Kooperationsprojekt. Dialog mit Bibliotheken 23(1): 37–40
4. Bennett R, Hengel-Dittrich C, O'Neill ET, Tillett BB (2007) Viaf (virtual international authority file): Linking the Deutsche Nationalbibliothek and Library of Congress name authority files. Int Cataloguing Bibliogr Control 36:12–18
5. Zhao F (2022) A systematic review of Wikidata in digital humanities projects. Digit Scholarsh Humanit 38(2):852–874. https://doi.org/10.1093/llc/fqac083
6. Güntsch A, Groom Q, Ernst M, Holetschek J, Plank A, Röpert D, Fichtmüller D, Shorthouse DP, Hyam R, Dillen M, Trekels M, Haston E, Rainer H (2021) A botanical demonstration of the potential of linking data using unique identifiers for people. PLoS ONE 16(12):1–11. https://doi.org/10.1371/journal.pone.0261130
7. Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen AE, Haston E (2020) People are essential to linking biodiversity data. Database. https://doi.org/10.1093/database/baaa072
8. Henrich A, Gradl T (2021) Integration von Forschungsdaten. Wie können Forschungsinfrastrukturen helfen? In: Seng E-M, Göttmann F (eds) Innovation in der Bauwirtschaft: Wesersandstein, 16th edn. De Gruyter, Berlin, Boston, pp 749–762 https://doi.org/10.1515/9783110538915-039
9. Gradl T, Henrich A (2016) Die DARIAH-DE-Föderationsarchitektur—Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen. Bibliothek Forsch Prax 40(2):222–228. https://doi.org/10.1515/bfp-2016-0027
10. Fruth L, Gradl T, Henrich A (2025) ADISS: Authority data integration search system. In: Datenbanksysteme für Business, Technologie und Web—Workshopband (BTW 2025). Gesellschaft für Informatik, Bonn, pp 189–200 https://doi.org/10.18420/BTW2025-123
11. Wiederhold RA, Reeve GF (2021) Authority control today: principles, practices, and trends. Cataloging Classif Q 59(2–3):129–158. https://doi.org/10.1080/01639374.2021.1881009
12. Beek W, Raad J, Wielemaker J, Harmelen F (2018) sameAs.cc: The closure of 500M owl:sameAs statements. In: Gangemi A, Nav-

igli R, Vidal M-E, Hitzler P, Troncy R, Hollink L, Tordai A, Alam M (eds) The Semantic Web. Springer, Cham, pp 65–80 https://doi.org/10.1007/978-3-319-93417-4_5
13. Adams B (2021) Chronotopic information interaction: Integrating temporal and spatial structure for historical indexing and interactive search. Digit Scholarsh Humanit 36(3):525–541. https://doi.org/10.1093/llc/fqaa049
14. Koch I, Ribeiro C, Poveda-Villalón M, Rico M, Teixeira Lopes C (2024) Enriching archival linked data descriptions with information from Wikidata and DBpedia. In: Antonacopoulos A, Hinze A, Piwowarski B, Coustaty M, Di Nunzio GM, Gelati F, Vanderschantz N (eds) Linking Theory and Practice of Digital Libraries. Springer, Cham, pp 396–412 https://doi.org/10.1007/978-3-031-72437-4_23
15. Erlinger C (2019) Sächsische Ortsdaten in der linked open data cloud: Teilautomatisierte Anreicherung und Analyse der HOV-ID in Wikidata. https://saxorum.hypotheses.org/2917. Accessed 2025-04-16. https://doi.org/10.58079/tw8g
16. Myntti J, Lewis N, McCormack AM, Rockwell K (2020) Regional connections to national authority files. Cataloging Classif Q 58(1):76–89. https://doi.org/10.1080/01639374.2019.1690087
17. Ravelli E, Mataloni MC (2022) Integrated search system: Evolving the authority files. JLISit 13(1):335–346. https://doi.org/10.4403/jlis.it-12716
18. Angjeli A, Ewan MA, Boulet V (2014) ISNI and VIAF—transforming ways of trustfully consolidating identities. In: IFLA WLIC 2014—Lyon—Libraries, Citizens, Societies: Confluence for Knowledge (226). Lyon, France, pp 16–22
19. Hickey TB, Toves JA (2014) Managing ambiguity in VIAF. D-Lib Mag. https://doi.org/10.1045/july2014-hickey
20. Toves JA, Hickey TB (2014) Parsing and matching dates in VIAF. Code4lib J 26
21. Bianchini C, Bargioni S, Girolamo CCPDS (2021) Beyond VIAF: Wikidata as a complementary tool for authority control in libraries. Inf Technol Libr. https://doi.org/10.6017/ital.v40i2.12959
22. Menzel S, Schnaitter H, Zinck J, Petras V, Neudecker C, Labusch K, Leitner E, Rehm G (2021) Named entity linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten. In: Franke-Maier M, Kasprzik A, Ledl A, Schürmann H (eds) Qualität in der Inhaltserschließung, vol 2021. De Gruyter Saur, Berlin, Boston, pp 229–258 https://doi.org/10.1515/9783110691597-012
23. Jegan R, Fruth L, Gradl T, Henrich A (2023) Integrating access to authority data for improved interoperability of research data in the digital humanities. In: Datenbanksysteme für Business, Technologie und Web (BTW). LNI, P-331 edn. Gesellschaft für Informatik e.V., Bonn, pp 829–836 https://doi.org/10.18420/BTW2023-54
24. Karam N, Lorenz RH, Müller-Birn C (2017) The GFbio terminology service—enabling a research data management beyond data heterogeneity. Freie Universität Berlin, Berlin https://doi.org/10.17169/refubium-25113
25. Golden P, Shaw R (2015) Period assertion as nanopublication: the PeriodO period gazetteer. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15 companion. Association for Computing Machinery, New York, NY, USA, pp 1013–1018 https://doi.org/10.1145/2740908.2742021
26. Brase J (2009) DataCite—a global registration agency for research data. In: Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, pp 257–261 https://doi.org/10.1109/COINFO.2009.66