ADISS: Authority Data Integration Search System

Leon Fruth [1], Tobias Gradl [1], and Andreas Henrich [1]

Abstract: This paper introduces ADISS, a generic search system designed to integrate heterogeneous authority file providers. Authority data is used to unambiguously identify entities such as persons, places, and organizations. As individual data providers usually do not offer both broad, universal data and domain-specific, in-depth data at the same time, in many application scenarios a combined access to multiple datasets is required to support real-world use cases. In the context of Digital Humanities this combination of multiple authority data providers improves the resolution of ambiguities in data curation processes. Our work in this direction is mainly motivated by two projects that require semi-automatic retrieval, as well as user-centered search scenarios for different authority file providers. Instead of using multiple existing endpoints to access the various datasets, we gather the heterogeneous data and make it accessible via integrated query and result models, thus simplifying access at the level of technical interfaces and schema integration. In this paper, we present our highly configurable search API, which offers a diverse range of search and filtering options. We show that by its generic and highly configurable nature, our system is adaptable and reusable for a diverse set of use cases and conclude the paper with ideas for further steps and 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 [Wi16]. Authority files play a key role regarding the implementation of the FAIR principles by enabling unambiguous identification of entities like locations, organizations or persons. Rather than relying solely on textual descriptions, links to entity representations in knowledge bases enrich and contextualize research data and thus enhance their overall interoperability, making it easier to discover and use [Ha21].

There are several 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 categorized into three groups. The first group consists of highly regulated authority file providers such as the Gemeinsame Normdatei (GND)[2] or the Virtual International Authority File (VIAF) that are managed by trusted institutions and libraries. The GND is a project of the German National

[1] University of Bamberg, Media Informatics Group, An der Weberei 5, 96047 Bamberg, Germany,
leon.fruth@uni-bamberg.de, https://orcid.org/0009-0001-2128-3025;
tobias.gradl@uni-bamberg.de, https://orcid.org/0000-0002-1392-2464;
andreas.henrich@uni-bamberg.de, https://orcid.org/0000-0002-5074-3254

[2] https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html All links accessed on 23-12-2024

doi:10.18420/BTW2025-123

Library (DNB) and is strictly maintained within quality standards and legal settings [BP11]. 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 [Be06]. Due to their focus on quality and institutionally induced biases[3], providers like the GND expose sets of highly regulated and curated entries that are naturally limited in quantity.

The second group includes community-based projects such as Wikidata[4] which hold substantial amounts of user-generated data. Wikidata is commonly utilized in Digital Humanities (DH) projects to support metadata curation, annotation and Named Entity Recognition (NER) [Zh22]. Additionally, Wikidata is also utilized in natural sciences to improve data integration and accessibility by identifying people that are associated with entities, such as specimens and taxonomic names [Gr20; Gü21]. Despite unmatched degrees of quantity, these community-based platforms pose challenges related to data accuracy and exhibit biases driven by public interests. [Zh22]

Lastly, the third group consists of small, specialized data repositories. For instance, Memorial Archives[5] contains entity descriptions specific to the Flossenbürg concentration camp, while the historic place names in Bavaria from Geschichte Bayerns[6] can be a valuable resource for certain research communities. Often being very detailed in their respective contextual setting, specialized data repositories attempt to fill specific gaps, but do not contain descriptions of more general entities.

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 curation possibilities – enhancing the quality of the structure, detail, and semantic richness of the data and thus its findability [HG21]. Such requirements on integrated access to several authority file providers for two DH projects will be elaborated in the next section.

Thereafter, this paper introduces 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 component is based on preliminary work in the infrastructural contexts of DARIAH-DE and CLARIAH-DE [GH16].

3 National and cultural biases in the case of national libraries due to their mission
4 https://www.wikidata.org/
5 https://memorial-archives.international/
6 https://www.geschichte-bayerns.de/ortsnamen

2 Motivation

The idea of consolidating access to multiple authority file providers results from the informational needs in the research infrastructures Text+[7] and Oral-History.Digital (oh.d)[8]. 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, organizations 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 NER 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 generic 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. 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[9], 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 contextualization 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, organizations 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 a authority file, then the resolved entities 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.

[7] https://text-plus.org/
[8] https://www.oral-history.digital/
[9] http://registry.text-plus.org

Both Text+ and oh.d implement functionalities to contextualize 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 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 above usage scenarios. 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, OpenStreetMap (OSM) and Geonames.

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 close to 10 million total entities (as of 18.12.2024). The data can be obtained as a data dump[10] or queried via a web-API[11]. 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. However, other authority data sources can be used to fill these gaps for entities that exist in multiple sources and contain at least an 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. Due to their community-based nature, Wikidata and other providers have some

[10] https://data.dnb.de/opendata/
[11] https://lobid.org/gnd/api

quality issues, such as (near) duplicate entries or incorrect links to other authority data providers [Me21]. The data can be accessed through dumps in multiple formats[12] as well as via SPARQL-endpoint[13] or the Wikibase REST-API[14]. Still, these endpoints do not support options like fuzziness or geographical filtering at present.

An additional data provider utilized for the described use cases is OSM[15], 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[16], which offers text searches, but lacks functionalities such as fuzziness and geographical filtering. Another option can be found in Photon[17], a service that allows some geographical prioritizing, but does not include many language labels that are present in Nominatim.

Geonames[18] 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.

4 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, ranking results from different APIs is challenging due to differences in data types, query languages and retrieval models between providers.

The second approach, which is presented 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 also accessing various data sources concurrently. This method provides arguably better performance than the previously mentioned approach as it reduces network latency by minimizing requests down to only one instead of N requests for each of the N considered data providers.

Moreover, heterogeneous data is integrated and cached into a homogeneous format during indexing rather than during runtime when the providers are queried, which further improves

12 https://dumps.wikimedia.org/wikidatawiki/entities/
13 https://query.wikidata.org/
14 https://www.mediawiki.org/wiki/Wikibase/API
15 https://www.openstreetmap.org
16 https://nominatim.org/
17 https://photon.komoot.io/
18 https://www.geonames.org/

the performance. Additionally, highly specialized data sources that lack an API can be added for research projects requiring such information.

For scalability and reduced maintenance efforts, a hybrid approach combining download and API-access could also be used, which has been presented in [Je23]. 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 [KLM17]. 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. However, there are currently no problems in this respect. The data is regularly updated to meet the needs of the above usage scenarios, and the combined storage requirement for the five integrated databases is 67 gigabytes.

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.

5 Implementation

This section presents central aspects of the implementation of ADISS. Figure 1 shows the architecture of the service. For each data provider, an individual data wrapper processes the respective format and structure. The first three steps in the data pipeline are considered in the first subsection 5.1. In the subsequent subsection 5.2, the concept for the integrated data model is outlined. In subsection 5.3 the Generic Search (GS) API is presented with its most relevant features and configuration options. This search component provides search and retrieval functionalities for the processed data.

5.1 Data Collection and Processing

For each data provider that is required in the application context and offers download options for their authority data, an individual data wrapper[19] is implemented using Java and the Spring Boot Framework. These wrappers share core functionalities, but are adapted to the respective structural constraints of the data. Additionally, existing functionality in libraries like osm2pgsql[20] for OSM data dumps or the Wikidata Toolkit[21] for Wikidata have been reused for data processing.

[19] https://gitlab.com/minfba/resinfra/adiss
[20] https://osm2pgsql.org/
[21] https://www.mediawiki.org/wiki/Wikidata_Toolkit

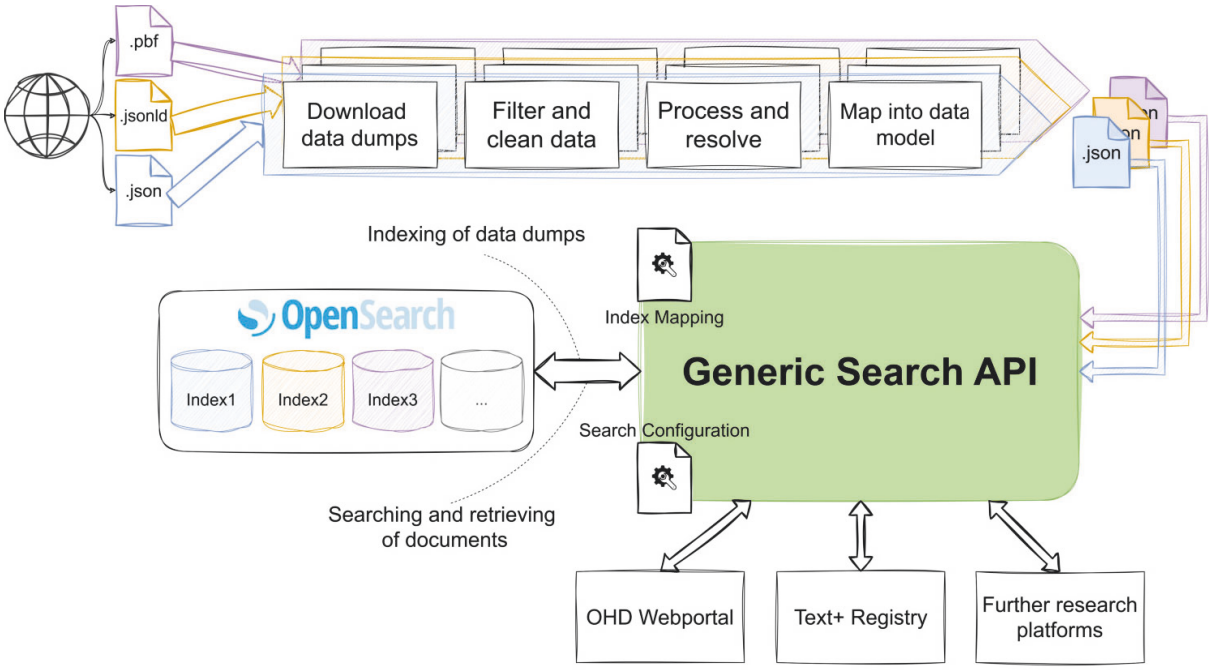

Fig. 1: The integrated authority data service.

The process starts with downloading the data dump from its respective source in the available format, GND entities in JSON-LD format, OSM data dumps in PBF format, and Wikidata in JSON or XML dumps. Next, the extracted data is filtered to include only entries of relevance for our use cases. In Wikidata, for instance, entities that are subclasses of irrelevant entries such as astronomical objects[22] or chemical compounds[23] are not considered. In addition, irrelevant attributions and media contents, are removed. With OSM, entries like highways or power lines are ignored.

Moreover, some fields are required to be structurally adapted; for example, coordinates which are represented in varying standards such as well-known text (WKT), are all transformed into the GeoJSON format. To facilitate comprehensive search capabilities on data from multiple sources, similar harmonization 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 also present in different ISO-formats.

5.2 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. This internal data model is composed of two main parts:

[22] https://www.wikidata.org/wiki/Q6999
[23] https://www.wikidata.org/wiki/Q11173

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 Listing 1.

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

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

Listing 1: Simplified example for the `integration` field of an entity.

As a final result of data preprocessing, data is saved in an intermediary JSON format. This explicit intermediary 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 can then be indexed with our API.

5.3 Search and Retrieval

The Generic Search (GS) API, utilized for indexing and retrieving the data, utilizes OpenSearch as backend.[24] 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 section 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 tweak different OpenSearch functionalities and provide them in terms of search profiles that can be tailored to the specific needs of use cases. A search query can be set up by combining different query types, such as full-text queries that support Lucene's Query Parser Syntax[25], as well as queries for (near-)exact matches in keyword fields. In order to support the varying search scenarios, described in section 2, different combinations of queries are used. 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 uses a ranking mostly based on full-text queries for high-recall results. Notably, additional configurations, such as aggregations, search suggestions, keyword highlighting, and the definition of geographical fields for utilizing distance or polygon filters can be added to the search profiles. In the user-centered search scenario for instance, keywords are aggregated to give the user an overview of the results, geographical fields are defined and query suggestions are presented.

Another important feature within the search configuration is the dynamic alteration of the result relevance score using scripts. This is currently employed to favor entities that contain links to matching entities of other providers, which we interpret as an indicator of the general relevance. These entries are preferred over others with a similar score, but cannot be potentially merged with other entities. 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 specialized 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 customizing a search request can be specified in terms of the `POST` request body[26]. An example is given in listing 2. The request considers all

[24] https://gitlab.com/minfba/resinfra/generic-search/generic-search-api
[25] https://lucene.apache.org/core/2_9_4/queryparsersyntax.html
[26] https://c102-142.cloud.gwdg.de/adiss-gs/search/default

indices that contain geographical data (line 3). It looks for entities that contain a polygon of coordinates (line 4), and are within `10 km` of some coordinates (line 5-6). The linked entities of the given results are joined in the result list (line 7).

```
1  {
2    "query": "Bamberg",
3    "indices": ["geo"],
4    "filter": {"exists": "_integration.coordinates",
5               "_integration.countryCode": "DE"},
6    "coordinates": [49.9031, 10.8695],
7    "distance": 10000,
8    "joinLinks": true
9  }
```

Listing 2: JSON body for a search request.

6 Conclusion and Future Work

This work introduced ADISS, a search system designed to integrate multiple authority file providers. The system enables the simultaneous querying of data sources to retrieve a wide range of enriched authority data for data curation processes. Motivated by use cases from two research projects in the field of DH that currently utilize the API, we have demonstrated how our generic and configurable approach facilitates diverse methods of accessing this data.

The development of this system is ongoing, with plans to implement additional features, some of which have been briefly described in the preceding sections. Prospectively, more data providers will be integrated. Beyond the data presented in this paper, further data sources will be included such as Memorial Archives and historic place names from Geschichte Bayerns. Person records e.g. from ORCID[27] would extend the set of person entities of the GND and Wikidata, which we already started experiencing with. Overall, persons and other entity types will receive more focus in the future due to their relevance for our primary use cases.

To improve the relevance scoring of the search results, it may be advantageous to consider the usages and references of individual entities within research datasets. Entities utilized for enriching research metadata should receive higher consideration due to their employment frequency. However, any optimization might lead to unwanted side-effects like hiding entities from highly specialized providers. Due to the configurability of our solution, we expect to be able to mitigate such risks by classifying information needs and addressing them with dedicated search profiles. The introduction and fine-tuning of various search configurations

[27] https://orcid.org/

for specific use cases, and leveraging the flexibility of the system further enhances the search effectiveness. First experiences in this respect, particularly from the population of the Text+ Registry[28], are encouraging and need to be validated in comprehensive systematic evaluations in the future.

Moreover, we are currently investigating into standards, such as the Schema.org Place schema[29] or the Datacite schema[30], that can be applied to our integration model to enhance the interoperability of our system.

References

[Be06] Bennett, R.; Hengel-Dittrich, C.; O'Neill, E. T.; Tillett, B. B.: Viaf (virtual international authority file): Linking die deutsche bibliothek and library of congress name authority files. In: World library and information congress: 72nd IFLA general conference and council. 2006.

[BP11] Behrens-Neumann, R.; Pfeifer, B.: Die Gemeinsame Normdatei: Ein Kooperationsprojekt. German, Dialog mit Bibliotheken 23 (1), ed. by Nationalbibliothek, D., pp. 37–40, 2011, ISSN: 0936-1138, URL: https://katalog.fid-bbi.de/Record/185-11152.

[GH16] Gradl, T.; Henrich, A. Bibliothek Forschung und Praxis 40 (2), pp. 222–228, 2016, DOI: doi:10.1515/bfp-2016-0027, URL: https://doi.org/10.1515/bfp-2016-0027.

[Gr20] Groom, Q.; Güntsch, A.; Huybrechts, P.; Kearney, N.; Leachman, S.; Nicolson, N.; Page, R. D. M.; Shorthouse, D. P.; Thessen, A. E.; Haston, E.: People are essential to linking biodiversity data. Database 2020, baaa072, 2020, ISSN: 1758-0463, DOI: 10.1093/database/baaa072, eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baaa072/36143659/baaa072.pdf, URL: https://doi.org/10.1093/database/baaa072.

[Gü21] Güntsch, A.; Groom, Q.; Ernst, M.; Holetschek, J.; Plank, A.; Röpert, D.; Fichtmüller, D.; Shorthouse, D. P.; Hyam, R.; Dillen, M.; Trekels, M.; Haston, E.; Rainer, H.: A botanical demonstration of the potential of linking data using unique identifiers for people. PLOS ONE 16 (12), pp. 1–11, 2021, DOI: 10.1371/journal.pone.0261130, URL: https://doi.org/10.1371/journal.pone.0261130.

[Ha21] Hawkins, A.: Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. Archival Science 22, pp. 319–344, 2021, DOI: doi:10.1007/s10502-021-09381-0, URL: https://doi.org/10.1007/s10502-021-09381-0.

[HG21] Henrich, A.; Gradl, T.: Integration von Forschungsdaten. Wie können Forschungsinfrastrukturen helfen? In (Seng, E.-M.; Göttmann, F., eds.): Wesersandstein vom 16. bis 19. JahrhundertWeser Sandstone from the 16th to the 19th Century Architektur und Digital HumanitiesArchitecture and Digital Humanities. De Gruyter, Berlin, Boston, pp. 749–762, 2021, ISBN: 9783110538915, DOI: doi:10.1515/9783110538915-039, URL: https://doi.org/10.1515/9783110538915-039.

[28] https://registry.text-plus.org/

[29] https://schema.org/Place

[30] https://schema.datacite.org/meta/kernel-4.6/

[Je23] Jegan, R.; Fruth, L.; Gradl, T.; Henrich, A.: Integrating Access to Authority Data for Improved Interoperability of Research Data in the Digital Humanities. In (König-Ries, B.; Scherzinger, S.; Lehner, W.; Vossen, G., eds.): Datenbanksysteme für Business, Technologie und Web (BTW 2023), 20. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme"(DBIS), 06.-10, März 2023, Dresden, Germany, Proceedings. Vol. P-331. LNI, Gesellschaft für Informatik e.V., pp. 829–836, 2023, DOI: 10.18420/BTW2023-54, URL: https://doi.org/10.18420/BTW2023-54.

[KLM17] Karam, N.; Lorenz, R. H.; Müller-Birn, C.: The GFBio Terminology Service. 2017.

[Me21] Menzel, S.; Schnaitter, H.; Zinck, J.; Petras, V.; Neudecker, C.; Labusch, K.; Leitner, E.; Rehm, G.: 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. De Gruyter Saur, Berlin, Boston, pp. 229–258, 2021, ISBN: 9783110691597, DOI: doi:10.1515/9783110691597-012, URL: https://doi.org/10.1515/9783110691597-012.

[Wi16] Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1), p. 160018, 2016, ISSN: 2052-4463, DOI: 10.1038/sdata.2016.18, URL: https://doi.org/10.1038/sdata.2016.18.

[Zh22] Zhao, F.: A systematic review of Wikidata in Digital Humanities projects. Digital Scholarship in the Humanities 38 (2), pp. 852–874, 2022, ISSN: 2055-7671, DOI: 10.1093/llc/fqac083, eprint: https://academic.oup.com/dsh/article-pdf/38/2/852/50488385/fqac083.pdf, URL: https://doi.org/10.1093/llc/fqac083.