Integrating Access to Authority Data for Improved Interoperability of Research Data in the Digital Humanities

Robin Jegan[1] Leon Fruth[1] Tobias Gradl[1] Andreas Henrich[1]

Abstract: Authority data is used to unambiguously identify persons, organizations and places. In this paper, a means to integrate access to several providers of authority data into data curation processes is described, which facilitates disambiguation of geographic data. Combined access to general datasets, in our case the Integrated Authority File (GND), as well as highly specialized datasets, here the Memorial Archives, improves the resolution of ambiguities and particularly benefits use cases of the Digital Humanities. The integration is necessary in order to abstract from technical, syntactical and semantic heterogeneity of the providers. Operations such as querying geographic information and receiving enriched data from different data sources are facilitated. An overview of the goals of the system, related projects and authority data providers are presented, as well as details on the implementation and further steps.

Keywords: authority file; geographic databases; data integration

1 Introduction

FAIR[2] has become a prominent keyword in academia that summarizes fundamental requirements of research data management. Authority files play a key role with respect to multiple FAIR principles in that entities such as persons, organizations and places can be unambiguously identified. Potentially ambiguous textual descriptions of entities (i.e. name attributions) can thus be replaced with references to their representation in authority files – improving the interoperability and in consequence the findability and reusability of data.

Numerous sources of authority data exist and provide access in terms of distinct data structures and formats – either by means of downloadable archives or in the form of accessible APIs. Data selection and their semantic and structural representation are influenced by requirements and contexts of respective providers. Authority files of national libraries, such as the Integrated Authority File (Gemeinsame Normdatei, (GND)[3] of the German National Library (DNB) aggregate authority data within their legal setting[4] and thus with a national bias. In contrast, universal data sources such as GeoNames[5] might expose a

[1] Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik, An der Weberei 5, 96047 Bamberg, Germany, [robin.jegan,leon.fruth,tobias.gradl,andreas.henrich]@uni-bamberg.de

[2] Findability, Accessibility, Interoperability, and Reuse of digital assets https://www.go-fair.org/fair-principles/ All links accessed on 18-01-2023.

[3] https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

[4] https://www.gesetze-im-internet.de/dnbg/__2.html

[5] https://www.geonames.org/

doi:10.18420/BTW2023-54

reduced descriptive depth. As an example of a highly specific normative data source, the Memorial Archives[6] contain entity descriptions in the specific context of the Flossenbürg Concentration Camp and as such, entities that are often unavailable in more generic authority files.

Concurrent access to multiple authority files is required if information of universal and specific files need to be referenced and combined. Use cases with such integrative requirements are often found in the Digital Humanities (DH), where historic entities and attributes such as names and chronologies are of particular relevance. As an example, the project Oral-History.Digital (OH.D)[7] develops a curation and research platform for collections of audiovisually recorded narrative interviews. Metadata and transcripts of interviews usually include geographic data such as historic places connected to the interviewed person or the location the interview was recorded. In order to clearly label and identify the spatial information, authority file providers are used to differentiate between ambiguous places. As such, interviews with survivors of the Flossenbürg concentration camp might contain references to specific places such as satellite camps, which can be found in the Memorial Archives. On the other hand, that same interview might reference commonly known places such as cities and buildings, that can be referenced by using the GND.

In this paper, we present an idea and software component that mediates between the required contextualization of research data (such as interviews) and the wide-spread authority data that is available. After presenting a more detailed perspective on data providers and the technological complexities of accessing provided data, we introduce a service that mitigates these complexities by modeling and mapping between heterogeneous data structures, providing integrated access and presenting authority data in an integrated manner. For an initial realization of the service, we have focused on geographical authority data, but expect to integrate other entity types in the near future.

Our contributions are twofold. After analyzing and evaluating data providers and use cases, we identify possibilities to enable access to heterogeneous spatial data through dedicated APIs. Furthermore, we propose integrative access to spatial data by means of a Transformation Service, which is realized on the basis of existing modeling and mapping capabilities, see 4.2.

Authority data in our paper includes data from authority file providers such as the GND, which are curated by their holding institutions, as well as community-based projects like WikiData[8]. With respect to our user-case, both types of data serve the same purpose, identifying entities, and from a technical perspective can be handled similarly in order to achieve the objectives presented in this paper.

[6] https://memorial-archives.international/
[7] https://www.oral-history.digital/
[8] https://www.wikidata.org/

2 Related Work

The features of authority data include identifiers that can serve to clearly identify entities. In contrast to curated qualitative data from authority file providers, community-based projects, e.g. WikiData or Nominatim[9], are characterized by user-created data and their immense volume, and can therefore be used to interconnect with information from providers like the GND, to enrich and broaden the data.

Regarding the flexible integration of different authority file providers for spatial data, to the best of our knowledge, no solution is available. A similar, but more general effort has been made to create an overarching authority file across multiple national libraries and archives, resulting in the Virtual International Authority File (VIAF) [Be06]. Other datasets such as the Library of Congress Name Authority Files (LCNAF)[10] serve the same purpose. For these resources, VIAF and LCNAF, an OpenRefine implementation[11] connecting the two datasets has been implemented. The Open Researcher and Contribution Identifier (ORCID), a database for researchers, is another authority file, mainly including data on persons and their publications. However, ORCID has to be separated from VIAF and LCNAF, since researchers themselves can edit, or even remove, their entries and thus the persistence of ORCID data cannot be assured [Pi22, p. 137-138].

On a national level, there have been efforts to integrate authority files from many independent archives and libraries, such as the national network of Italian libraries (Servizio Bibliotecario Nazionale), whose goal is to integrate roughly 12,000 libraries across Italy and their authority files [Ma22]. Another project describes the efforts in the research field technology assessment and its portal for specialists, in which authority files and other data sources are used to identify persons, organization and other data [Ho18]. There, the GND is used in combination with data from ORCID or WikiData in order to identify entities.

Regarding spatial data, the Geobrowser project of the Digital Research Infrastructure for the Arts and Humanities (DARIAH) enables the upload of spatial data in various file formats and visualization on a geographic map [12]. The integration of geographic data in KML, KMZ or CSV formats is available here as well as various options regarding the presentation of the data, e.g. individual layers via file upload (again in KML or KMZ file formats), as well as ArcGIS layers or predefined layers such as historical data for the Roman Empire [Ko16].

The requirement to integrate different authority data providers for geographic data in one infrastructure and thus to enable querying on this heterogeneous data is a new application scenario and will be presented below, after a closer look on data providers.

[9] https://nominatim.org/
[10] https://authorities.loc.gov/
[11] https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation/
[12] https://geobrowser.de.dariah.eu/

3 Data Providers

During an initial requirements analysis in the OH.D project, several providers for authority data were identified. These are divided into generic and specialized data sources.

A generic authority file provider that was used for this implementation is the GND entity Geografikum. The access to the data is enabled through the data service entity facts, which comprises a web interface as well as dataset dumps[13] including over 322,000 GND entities. The dataset mostly consists of geographical places, including information like different names of the place and the geographical area code. The metadata included in this dataset is however lacking in some aspects. Many entities contain little to no alternative names, like language or historical variants, which can help to resolve ambiguities. Furthermore, geographic coordinates are only included in roughly 19% of all entities. Some entities contain references to other data sources, such as WikiData that can be used to further enrich the available data, for example to add historical places names.

Other generic authority file sources include GeoNames, a geographical database containing over 12 million entries, which can be downloaded as a data dump. Nominatim, the backbone of OpenStreetMap (OSM), provides an API, that allows access to over 7 billion OSM nodes. In addition to the online API, OSM data can be downloaded and has thus been reused in multiple projects. One prominent example can be found in Photon[14], an open-source search platform for OSM data.

Apart from the previously mentioned services, which offer generic spatial information, more specialized authority file providers are of particular relevance to the DH and thus represent interesting use-cases for this project. Interviews in OH.D, which comprise reports of contemporary witnesses, often include historic place names. Therefore, general authority data providers, even if they contain alternative place names in their datasets, are not suitable for connecting these historical places to their use in the interviews. In such cases, providers like the Memorial Archives, initiated by the Flossenbürg Concentration Camp Memorial[15], are needed for more specialized authority data. The Memorial Archives contain data on over 890,000 persons, 4,700 publications, 5,500 places and more. Furthermore, it incorporates references between those different data entries. Regarding spatial data, the type of the place, the time and name are included as well as coordinates. Here, an exemplary use-case for a historical researcher would be using the Memorial Archives in order to get all names for the town Chrastava in the Czech Republic. During the Second World War, a concentration camp near the city was also known by its Polish as well as German name, since the city is close to the Polish and German border. Through specialized data archives such as Memorial Archives, the names of this concentration camp in all three languages are available for further research[16].

[13] https://data.dnb.de/opendata/

[14] https://github.com/komoot/photon/

[15] https://www.gedenkstaette-flossenbuerg.de/

[16] https://memorial-archives.international/entities/show/56dc6ea9759c022fd48d5286

The Heterogeneity inherent in these data providers is apparent in different aspects. Direct API access is seldom available, which is why usually data dumps are downloaded and indexed. The dumps themselves come in various formats and require processing in a case-by-case manner.

4 Implementation

Authority data providers are heterogeneous in terms of access types (i.e. data downloads or query interfaces), models (i.e. query languages and schemata) and contexts (i.e. universal, national, specific). Our objective is to abstract from these aspects of heterogeneity and to facilitate integrative access to authority data. With regard to harmonizing access to authority data sources, section 4.1 describes the process and technical complexity of creating query APIs based on downloadable data dumps. With harmonized accessibility by means of existing or created APIs, section 4.2 focuses on the idea of overcoming model and context heterogeneity by introduction of the Transformation Service – a mediating component that translates between integrative models and the local models of a dynamic set of data sources.

4.1 Data Preparation & Indexing

To create accessible APIs based on available data dumps, microservices are created that process and index the different data sources in Elasticsearch[17] indices. These indices can be searched using the names of places and further filtered with information like country codes and spatial data. Additionally, individual entities can be requested by their respective identifier. The APIs use POST and GET Requests, with JSON request bodys and are automatically built and deployed as Docker images using continuous integration and deployment pipelines. First, the data dumps need to be processed and indexed. To keep data up to date they are reindexed in fixed intervals.

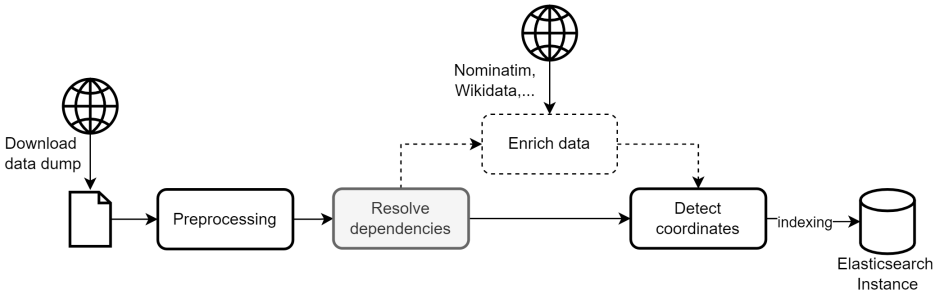

Fig. 1: Procedure of getting raw data, processing and indexing it to a Elasticsearch instance.

[17] https://www.elastic.co/elasticsearch/

Fig. 1 shows the process from downloading data sources to indexing them with Elasticsearch. The step *Resolve dependencies* is not necessary for all data sources. Also, the step *Enrich data* is not implemented yet and is further discussed in Section 5. So far the database dumps from the authority file providers GND Geografikum and GeoNames have been considered and indexed to an Elasticsearch instance, a next step will be to integrate the crawled Memorial Archives geographic data. Elasticsearch has been chosen due to other software projects that are connected or will be linked to this service in the future. It allows metadata to be queried in different ways, for example by filtering based on spatial information, such as coordinates or country codes.

For illustration purposes, the indexing of the Geografikum data dump is further described. The files containing the spatial data are available in several formats. Due to the use of Elasticsearch the format JSON-LD was chosen, since it is easy to integrate. The JSON-LD file contains entries with three different types of identifiers. One describes a GND entity (e.g. `https://d-nb.info/gnd/4004391-5`), which references entries with other types of identifiers, if available. The second type of identifier (e.g. `https://d-nb.info/gnd/4004391-5/about/`) carries further information about the entity. For almost one fifth of the GND entities, a third type of identifier (e.g. `_:node1gf2tc0d5x21379656`) is referenced, which points to an entry in the file containing the coordinates of the location. Before indexing, the dependencies between the entries in the JSON-LD file need to be resolved and integrated into one entity. Lastly, the coordinates are defined as so called Geopoints to allow querying and filtering by distance or location using Elasticsearch.

4.2 Data and Service Consolidation

Based on applied schemata and technical contexts of providers, APIs expect requests in specific forms (i.e. input models) and provide responses in terms of output models. In consequence, services with a need for integrated access to multiple data sources such as OH.D need to implement access to heterogeneous interfaces, input and output models. Emerging from the modeling capabilities of the DARIAH-DE Data Modeling Environment (DME)[18] [GH16, HG21] we propose the concept of a generic Transformation Service. Among other roles, the service acts as intermediary to online interfaces – mediating between integrative data demands and the accessible, yet heterogeneous sources of relevant data. Integrated query and result models are tailored to individual needs – here the OH.D portal and its technical setting.

Fig. 2 outlines the main functionality of the service with particular focus on the idea of interface mediation: Users of the OH.D portal formulate queries in terms of a defined query model and submit them to an API provided by the service. By applying mappings between integrated and local input models, queries can be translated and executed against applicable

[18] `https://de.dariah.eu/en/dme/`

data sources. Returned results are collected and sent to the user in terms of an integrated result model – again by applying relevant mappings.

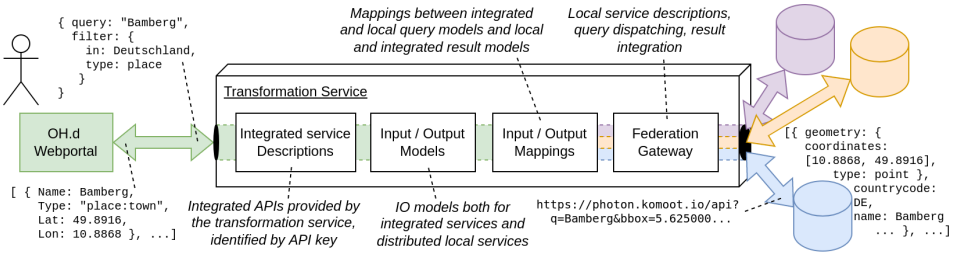

Fig. 2: Transformation service

In order to address initial use cases of OH.D, multiple APIs have been configured to each query one primary data source, whose results are enriched by queries to secondary sources. Primary and secondary geographical data sources can be configured individually per interview collection. All APIs oriented towards the OH.D portal share the same input and output models, despite binding to different data sources.

5 Discussion and Outlook

The advancements presented in this paper are embedded in the infrastructure of the language- and text-based infrastructure NFDI[19] project Text+ and can thus be re-used for future application scenarios. The integrated access to geographic information will enhance the capabilities of the Transformation Service (detailed in 4.2), as intermediary between requests, interfaces and data sources.

Still, the integration of data providers is affected by different file formats and access modalities. Thus, a procedural approach is necessary in order to handle the integration of additional data providers to enable comprehensive access.

To further improve the metadata, data providers described in Section 3 can be used. First, the number of geographic locations can be increased by utilizing more data providers like Nominatim and the Memorial Archives. Next, the data quality can be enriched by incorporating more historical place names or adding missing geographic information, like coordinates or country codes. Moreover, other types of entities such as persons or historical events should be considered in the future. Additional data dumps from GND and further sources of authority data can be used for this to index and search the data in a similar way to geographical data addressed in this work.

Our system is currently used as part of the OH.D project in order to identify spatial information and enrich metadata of interview transcripts. In doing so, feedback from researchers is raised and influences further development.

[19] https://www.nfdi.de/

Bibliography

[Be06] Bennett, Rick; Hengel-Dittrich, Christina; O'Neill, Edward T; Tillett, Barbara 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.

[GH16] Gradl, Tobias; Henrich, Andreas: Data Integration for the Arts and Humanities : A Language Theoretical Concept. In: Research and Advanced Technology for Digital Libraries 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5–9, 2016, Proceedings, pp. 281–293. Springer International Publishing, Cham, Switzerland, 2016.

[HG21] Henrich, Andreas; Gradl, Tobias: Integration von Forschungsdaten : Wie können Forschungsinfrastrukturen helfen? In: Innovation in der Bauwirtschaft, pp. 749–786. De Gruyter, Berlin, Boston, 2021.

[Ho18] Hommrich, Dirk; Pasucha, Beate; Razum, Matthias; Riehm, Ulrich: Normdaten und Datenanreicherung beim Fachportal openTA. Bibliotheksdienst, 52(3-4):248–265, 2018.

[Ko16] Kollatz, Thomas: Raum-zeit-analysen mit geo-browser und datasheet-editor. Bibliothek Forschung und Praxis, 40(2):229–233, 2016.

[Ma22] Mataloni, Maria Cristina: Integrated Search System: evolving the authority files. Bibliographic Control in the Digital Ecosystem, 7:335–346, 2022.

[Pi22] Piazzini, Tessa: Bibliographic control and institutional repositories: welcome to the jungle. Bibliographic Control in the Digital Ecosystem, 7:132–142, 2022.