Filling the gaps: Facilitating access to enriched research data

Recent years have seen a significant evolution in perspectives on research data and metadata in the domains of the Digital Humanities, driven by an emphasis on data management, the adoption of FAIR principles (Jacobsen et al. 2020), and increasing digital engagement in scholarly disciplines. This shift has been supported by the establishment of research data repositories, data aggregators and research infrastructures. Prominent services that provide integrated access to scholarly research data of the European region include Gams[1], Zenodo[2], and Europeana[3] as well as integrative services that are being developed within the four humanities-oriented consortia of National Research Data Infrastructure Germany (NFDI)[4]. Initiatives and platforms offer access to diverse data sets, each tailored to specific user groups, disciplinary domains, and connected data sources. Along with the common goals of increasing research data visibility and providing comprehensive access, another shared characteristic are the often stringent requirements these infrastructures impose on contributing repositories regarding data extent, quality control measures, and prescribed data formats.[5]

From the perspectives of data producers and providers – especially those with little or no dedicated funds for the installation and development of sophisticated repositories and interfaces – a critical question often remains unresolved: how to process, enrich, and transform existing data into formats compatible with aggregators and research infrastructures. A strategic design includes incorporating the requirements of targeted platforms for data export and utilizing export interfaces of repositories for transforming research data and metadata (see e. g. Frosini et al. 2018 and Conzett 2020). The complexity of these solutions varies, ranging from simple tasks like exporting standardized data from existing databases to more intricate procedures like creating XSLT[6] scripts for custom data processing. The complexity, and thus the propensity for errors, escalates in scenarios involving weakly structured data or data fragmented across distributed sources. Particularly challenging are cases where data cannot be conclusively migrated to context-specific target formats but requires ongoing operations in source systems and periodic exports.

Our Proposal

In response to these challenges, we propose a transformation service designed to aid data producers and users in navigating these difficulties. The service's primary functions include:

1. *Description of data models and mappings*: Users define source and target data representations, along with necessary enrichment and transformation steps. Focusing on the semantic aspects of data, these models are largely independent of technical specifics such as formats, protocols, and data access methods. The service

1 https://gams.uni-graz.at (the validity of all links in this text was last verified on December 12, 2023)
2 https://zenodo.org
3 https://www.europeana.eu
4 https://www.nfdi.de/consortia/?lang=en
5 see e.g. the Mapping Guidelines of the Europeana Data Model (EDM): https://pro.europeana.eu/page/edm-documentation
6 https://www.w3schools.com/xml/xsl_intro.asp

utilizes the Data Modeling Environment (DME, Henrich and Gradl 2021)[7], leveraging its capabilities for the generation and description of data models and mappings.

2. *Execution of Transformations*: The platform is conceptualized to execute specified transformation steps on live data on-demand. Data providers must specify an export interface, enabling the service to create a 'virtual dataset'. This dataset performs transformations and delivers data in the required specifications of the requesting platform. Caching mechanisms are implemented to ensure timely data delivery.
3. *Separation of Technology and Domain*: Utilizing the DME's functionality, the service maintains a separation between technological and logical aspects of data access, modeling, and transformation. This allows data providers to focus on their data expertise without the typical necessities to delve into technical details like format conversion or access protocols.

Basic scenarios: Virtual Datasets and Virtual APIs

The primary application of this service is to enrich data and unify fragmented datasets into cohesive, *virtual datasets*. This is especially beneficial for data providers with limited capabilities or resources. The service, as illustrated in figure 1, offers a threefold solution: 1) It allows for the specification and modeling of source data; 2) It provides the means to formulate rules for enriching, cleansing, and processing data; and 3) It enables the definition of desired output formats for the preprocessed data.

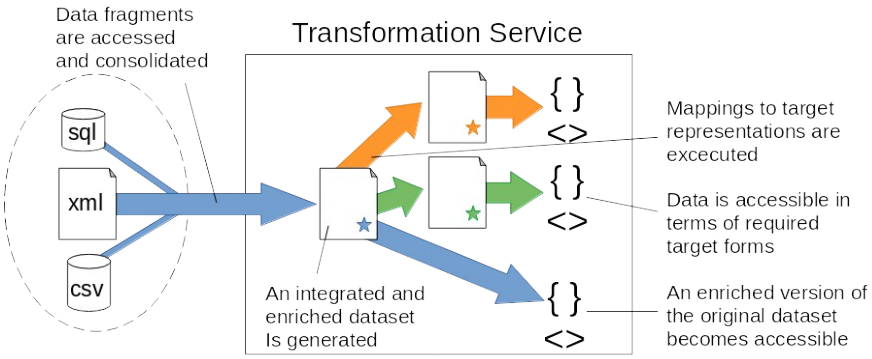


Figure 1: Virtual datasets provided by the transformation service

This approach allows the transformation of accessible online sources into REST-based APIs or OAI-PMH servers without altering the source system. The concept of virtual datasets is particularly useful for data intentionally created for website presentation. Based on the modeling capabilities of the DME, data can be extracted from websites and provided in terms of machine-readable interfaces without the need to implement any additional technical infrastructure.

The development of the transformation service also focuses on creating *virtual APIs*, which combine two transformation pipelines in order to support integrated request-response scenarios. Virtual APIs facilitate data conversion into specific input formats and then transform service-specific output formats into a uniform format for end users, ensuring a consistent and user-friendly API experience.

In initial testing, the virtual API concept has proven effective, especially for accessing

7 or less recent but in English language Gradl and Henrich (2017)

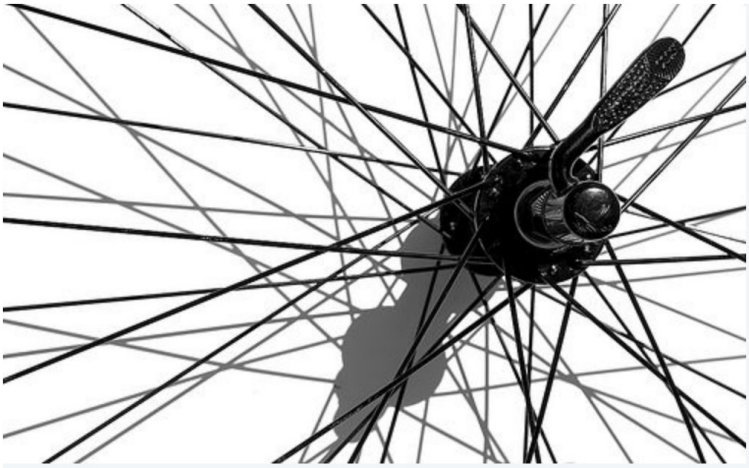
geographical reference data from sources like GeoNames, Wikidata, and OpenStreetMap (Jegan et al. 2023). Users can submit simple queries, which the virtual API then converts into the respective formats for each service. The results are standardized into a uniform format for user convenience.

To further illustrate the main idea of the transformation service, we present examples of how requirements were implemented during prototypical implementation and testing of the service. The examples are actual scenarios in the context of the CLARIAH-DE Tutorial Finder[8], the NFDI Text+ initiative[9] and the Oral.History-Digital project[10]. Due to the limited extent of this text, we mainly present results. Necessary models and the prototypical installations can be reached by following the links provided.

Example 1: Reusability of DH Tutorials

As part of the CLARIAH-DE project, a solution was sought that offers a comprehensive search in freely accessible and reusable teaching and training materials on research methods, procedures and tools in the field of digital humanities in various platforms and repositories. The CLARIAH-DE Tutorial Finder (Werthmann and Gradl 2022) provides access to heterogeneous sources of teaching and training materials in forms such as Markdown Documents in a Git Repository, the YouTube API or – as in the case of TeLeMaCo[11] – a static website (see figure 2). Despite the lack of a dedicated machine-usable interface, modeling capabilities of the DME allowed the specification of data extraction rules that emulate user navigation based on the keywords to individual tutorials, where metadata could be extracted from presented HTML tables[12]. As the result in figure 3 indicates, the transformation service implements a *virtual dataset* in the form of a REST API[13] that the CLARIAH-DE Tutorial Finder uses to harvest data of the collection.

8 https://teaching.clariah.de/search/
9 https://text-plus.org/en/
10 https://www.oral-history.digital/en/index.html
11 https://telemaco.clarin-d.uni-saarland.de/hub/browse/
12 https://teaching.clariah.de/dme/model/editor/5eb2c7b26117b123495e7f5d/
13 https://teaching.clariah.de/transformation/data/8DE2D94C116A8B5B02D4A9777A907E2A/4A1967A636451060460D38060B7C8759



The Linguistic Teaching Resources Hub

Image © Paul Watson, Licence CC BY-NC-SA 2.0

Home | What's new? | Browse | Login / Create Account | Advanced search | searc

Browse

Keywords

analysis of speech data, ANNIS, annotation, annotation management, annotation of speech data, AntConc, API, AQL, arch
Intelligence, audio and video transcription, authorship attribution, automatic annotation, automatic segmentation, bash,
Burrow's delta, character encoding, chi square, Chinese, chunking, CLARIN-D, CLAWS7, COCA, collocation, collocations, c
command line, Concordance, conversation analysis, corpus, corpus analysis, corpus annotation, corpus encoding, corpus
query, corpus search, corpus workbench, COSMAS II, counting, CQP, CQPweb, creation of speech data, data managment p
learning, DeReKo, Deutsch, DfR, DGD, DiaCollo, dialectology, dictionary, digital editing, digital humanities, DKPro, downlo

Figure 2: Static entry page of the TeLeMaCo site

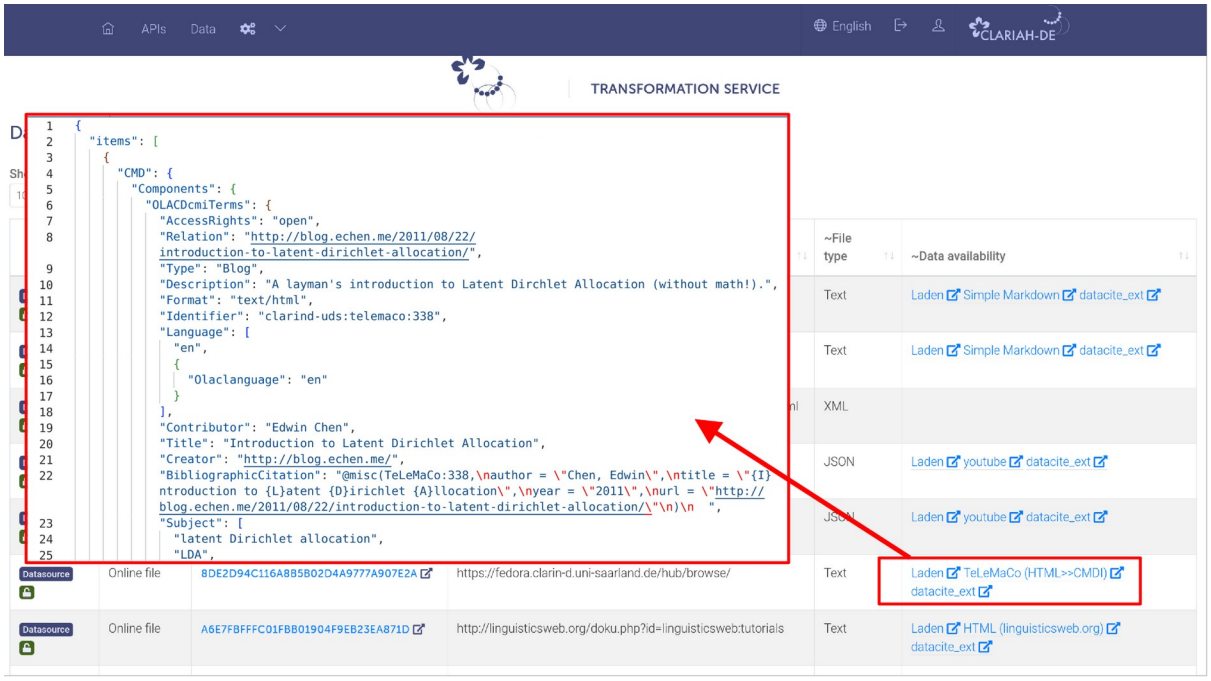


Figure 3: JSON Data that is produced by the transformation service

Example 2: Access to a Journal's TEI documents

A RSS feed allows *collection* access to the Zeitschrift für digitale Geisteswissenschaften

(ZfdG)[14] – as opposed to the individual access to TEI documents users gain when navigating the ZfdG website. Metadata available via the feed is limited to title, description, publication date and the link to the HTML page. URLs to the TEI documents of the articles are, however, accessible via buttons on the page. After obtaining the RSS feed, an HTML page can be called up by calling up the specified web address. By clicking on the buttons, then "Download XML", a representation of the article can be called up in TEI. In order to make all articles of the journal accessible in the form of TEI, the steps described are formalized within the framework of a data model in the DME.[15]

Again presenting data in terms of a *virtual dataset*, the transformation service facilitates access to the collection of TEI documents. The steps of users navigating through RSS feed, webpage to TEI document are modeled and the transformations service executes these steps to consume and finally provide all relevant data.

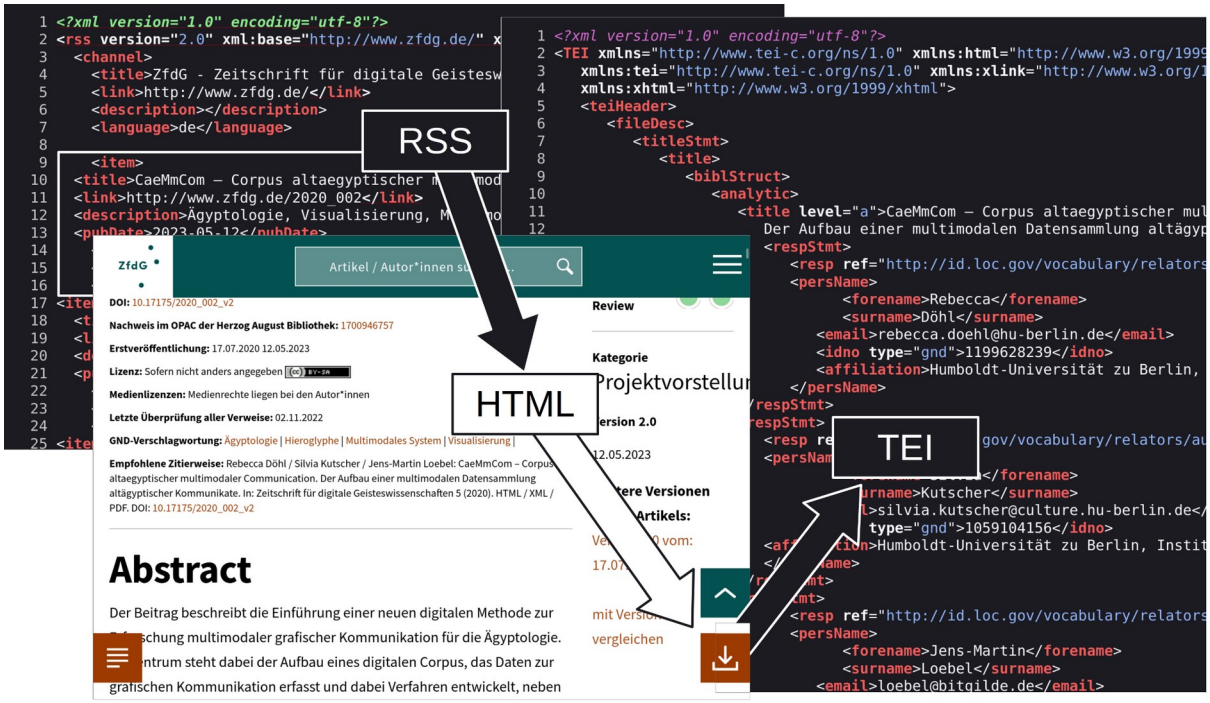


Figure 4: Steps for obtaining TEI documents based on the ZfdG RSS feed

Example 3: Interoperability of interfaces to authority data

APIs operate based on predefined models and the technical contexts of their respective providers. They are designed to receive requests in specific formats and to return responses according to designated output models. Consequently, services requiring integrated access to multiple data sources must accommodate a range of heterogeneous interfaces, input, and output models.

In this context, the transformation service functions as a *virtual, integrated API*.[16] It mediates between the requirements of integrative data access and the accessible, albeit diverse, data sources. This mediation is crucial in harmonizing the varying data structures and formats. Figure 5 delineates the primary functionality of this service, with an emphasis on the concept

14 https://zfdg.de/
15 https://dme.mww-forschung.de/dme/model/editor/59ca0aff06bffc019b193cbd/
16 https://c105-230.cloud.gwdg.de/transformation/endpoints/

of interface mediation.

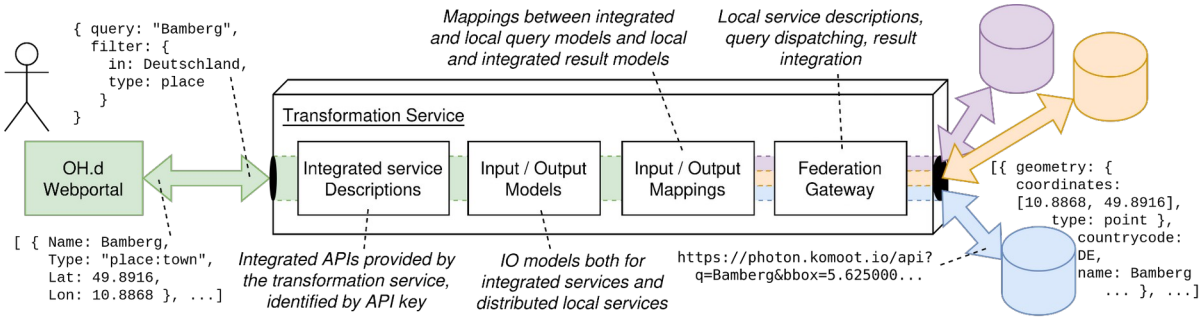


Figure 5: Virtual API for geographical authority data

Conclusion

Utilizing the data modeling and mapping capabilities of the DME, our transformation service improves the accessibility and interoperability of research data and metadata. Predominantly focusing on the development of virtual datasets and APIs, the service is adept at handling a diverse range of potential applications. It proves particularly beneficial in situations where data providers are limited by resource constraints, hindering their ability to migrate systems and data to architectures that comply with FAIR data principles.

The innovations presented in this paper have been incorporated into the NFDI Text+ infrastructure, which also ensures the adaptability and applicability of these developments in future projects and applications. Currently, the transformation service is undergoing evaluation in various contexts. This process is yielding valuable feedback from the academic community, which is instrumental in guiding the continuous refinement and evolution of the service.

References

Conzett, Philipp et al. (2020): “How to weave domain specific information sources into a large, FAIR data fabric for the Digital Humanities? The use of the Dataverse platform.“ https://doi.org/10.5281/zenodo.3879031

Frosini, Luca et al. (2018): “An Aggregation Framework for Digital Humanities Infrastructures: The PARTHENOS Experience“. SCIentific RESearch and Information Technology. Volume 8, Issue 1. http://dx.doi.org/10.2423/i22394303v8n1p33

Gradl, Tobias; Henrich, Andreas (2017): “Explicating knowledge on data models through domain specific languages“. Informatik 2017: 25.- 29. September 2017 Chemnitz, Germany, Proceedings. 1125–1136, https://doi.org/10.18420/in2017_114

Henrich, Andreas; Gradl, Tobias (2021): “Integration von Forschungsdaten : Wie können Forschungsinfrastrukturen helfen?”. Innovation in der Bauwirtschaft, pp. 749–786. De Gruyter, Berlin, Boston, https://doi.org/10.1515/9783110538915

Jacobsen et al. (2020): “FAIR Principles: Interpretations and Implementation Considerations”. Data Intelligence (2020) 2 (1-2): 10–29. https://doi.org/10.1162/dint_r_00024

Jegan, Robin et al. (2023): “Integrating Access to Authority Data for Improved

Interoperability of Research Data in the Digital Humanities“. Gesellschaft für Informatik e.V. https://doi.org/10.18420/BTW2023-54

Werthmann, Antonina; Gradl, Tobias (2022): “Der CLARIAH-DE Tutorial Finder. Eine Suchumgebung für Lehr- und Schulungsmaterialien in den Digital Humanities”. 8. Jahrestagung des Verbands. https://zenodo.org/doi/10.5281/zenodo.6304589