Secondary Publication

Gradl, Tobias; Kudella, Christoph; Lordick, Harald; Schulz, Daniela

Towards a Registry for Digital Resources : The Text+ Registry for Editions

Date of secondary publication: 12.12.2024

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-1054225

Primary publication

Gradl, Tobias; Kudella, Christoph; Lordick, Harald; Schulz, Daniela (2024): Towards a Registry for Digital Resources : The Text+ Registry for Editions, in: Datenbank-Spektrum : Zeitschrift für Datenbanktechnologie, Berlin ; Heidelberg: Springer, Vol. 24, Nr. 2, pp. 151–160, doi: 10.1007/s13222-024-00479-0.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

Datenbank-Spektrum (2024) 24:151–160
https://doi.org/10.1007/s13222-024-00479-0

SCHWERPUNKTBEITRAG

Towards a Registry for Digital Resources – The Text+ Registry for Editions

Tobias Gradl[1] · **Christoph Kudella**[2] · **Harald Lordick**[3] · **Daniela Schulz**[4]

Received: 19 February 2024 / Accepted: 8 May 2024 / Published online: 19 June 2024
© The Author(s) 2024

Abstract
The paper presents the Text+ Registry as a central system to describe and catalogue different kinds of resources. Using the example of editions and edition projects, desiderata and challenges regarding their discoverability are outlined, and the data model to cater for the wide range of different types of editions is described. With its overarching and flexible approach, the Text+ Registry is a versatile and flexible technical component, designed for future expansion and broad connectivity.

Keywords Metamodelling · Discoverability · Editions · Research Infrastructure · FAIR

1 Introduction

In the evolving landscape of scholarly research, the imperative of achieving FAIR (Findable, Accessible, Interoperable, and Reusable) resources has emerged as a cornerstone influencing the trajectory of knowledge generation, dissemination, and utilization [9]. As an initiative directed towards enhancing access to text and language resources, Text+[1] endeavors to construct a research data infrastructure that is both flexible and scalable, catering to the unique requirements of distinct academic disciplines, while also aspiring to serve as a central point of access across disciplines. Text+ initially focuses on three data domains and resources: *collections*, *lexical resources*, and *editions*, each playing a vital role across various disciplines within the Digital Humanities (DH). The consortium of Text+ comprises 26 data centers and 34 institutions, including research libraries, universities, and DH data centers, dedicated to provide data, expertise, tools, and services across disciplinary boundaries.

This paper delineates the architectural blueprint of the Text+ Registry as a pivotal component within the research infrastructure, addressing the exigent need for a centralized platform. This platform is envisaged not only to augment the visibility of scholarly resources but also to surmount longstanding barriers impeding the discoverability and accessibility of resources, adhering rigorously to the FAIR principles.

To explicate the intricacies and obstacles inherent in describing and discovering scholarly resources, Sect. 2 serves as a cornerstone, delving into editions as one of the Text+ data domains. This section outlines the identification of extant data sources, the conceptualization of a domain model, and the formalization of vocabularies and relations, tasks intrinsic to all Text+ data domains. Sect. 3 provides an exposition of our technical blueprint for the Registry, spotlighting the metamodel, user interface generation, and our layered approach to importing and consolidating existing catalogue data. Finally, Sect. 4 deliberates on implications

[1] https://text-plus.org/en/ (all online references in this document were last accessed on February 19, 2024).

✉ Tobias Gradl
tobias.gradl@uni-bamberg.de

Christoph Kudella
kudella@sub.uni-goettingen.de

Harald Lordick
lordick@steinheim-institut.org

Daniela Schulz
schulz@hab.de

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

[2] Göttingen State and University Library, University of Göttingen, Platz der Göttinger Sieben 1, 37073 Göttingen, Germany

[3] Salomon Ludwig Steinheim-Institute for German-Jewish History, associated institute of the University of Duisburg-Essen, Edmund-Körner-Platz 2, 45127 Essen, Germany

[4] Herzog August Bibliothek Wolfenbüttel, Lessingplatz 1, 38304 Wolfenbüttel, Germany

and outlines future avenues of work, thereby culminating this paper.

2 Discoverability of Editions

Scholarly editions constitute a key pillar within the Humanities and adjoining fields. According to Sahle, a scholarly edition can be defined as "critical representation of historic documents" [4]. The aim of scholarly editing is to prepare historical documents, texts or works in such a way (e.g. by adding annotations, apparatuses, commentaries, or also translations) that readers are provided with a reliable text to base their research on. Although traditional printed editions have long been prevalent and continue to persist, the proportion of digital editions and hybrid formats is steadily increasing. For instance, the Carl Maria von Weber Complete Edition (WeGA)[2] endeavors to compile all compositions, letters, diaries, and writings of Carl Maria von Weber. It involves meticulous research, the collation of manuscripts and sources, and the incorporation of digital technologies to make his oeuvre accessible both in both print and online formats, ensuring scholarly rigor and widespread availability.

2.1 Desiderata and Challenges

The undeniable relevance of editions for research notwithstanding, their discoverability presents a significant challenge deserving careful consideration. This issue arises within a complex context shaped by various factors: scholarly editions typically emerge within the framework of third-party-funded research projects, where not only the types of editions vary significantly, but also the funding bodies and formats used to disseminate results. Despite the enduring prevalence of the printed book as the primary output of scholarly editing, the landscape is evolving, with digital platforms increasingly complementing or even supplanting traditional formats. The nature of these projects being third-party funded leads to their inclusion in databases maintained by various research funding bodies, such as the Deutsche Forschungsgemeinschaft (DFG) and the Union of the German Academies of Sciences and Humanities (Akademienunion). However, within these databases, the emphasis remains primarily on aspects pertinent to research funding, such as grant identifiers, principal investigators, and subject classifications according to specific taxonomies of the individual funding bodies. Regrettably, explicit classification of projects as "editions" is frequently absent,[3] rendering it impossible to systematically retrieve the subset of data from these systems pertaining to this domain. Furthermore, these databases share a common limitation: they lack integration with external knowledge bases, such as authority data, which would enable more comprehensive linking of involved entities like individuals or institutions.

The prevalence of books as the primary output format of scholarly editing generally leads to editions being recorded in library or union catalogues respectively. However, locating editions within these catalogues systematically, despite their status as a distinct scholarly output, is notably deficient. For instance, systematically searching for a scholarly edition of letters may yield more effective results by querying the subject term "letter collections" rather than directly searching for "scholarly editions" and "letters". Moreover, the research outcomes of Digital Scholarly Editions encompass diverse formats, typically incorporating research data such as XML data according to the Text Encoding Initiative Guidelines (TEI-XML)[4] and digital facsimiles, which are quite often not catalogued at all.

Regarding the research data of editions themselves, two additional levels of documentation need to be considered. Firstly, there are the metadata with which the research data may have been enriched in repositories, which naturally vary from system to system in terms of the vocabularies used and the level of description provided. Secondly, there are the metadata that may be included within the research data themselves, primarily in the headers of TEI-XML-based editions, where similar heterogeneous descriptions may be encountered.

In addition to these sources, there are other directories, such as lists of editions maintained by the Specialized Information Services (SIS, Fachinformationsdienste)[5], or directories dedicated to digital editions curated by individuals or small groups, such as a cataloge of: Digital Scholarly Editions (DSE)[6] or the Catalogue of Digital Editions (CDE)[7].

This brief overview, by no means striving for completeness, seeks to illustrate the notably greater difficulty researchers encounter in pinpointing editions pertinent to their interests than one might initially presume. Not only do researchers need to manually query numerous systems, but they also need to intellectually synthesize various pieces of

[2] https://weber-gesamtausgabe.de/en/Project/Project_Description.html.

[3] The database system AGATE (https://agate.academy) represents a rare exception in this regard.

[4] https://tei-c.org/guidelines/.

[5] SISs are funded by the DFG and aim at providing scholars with appropriate specialised literature, information, and services that are relevant to their research.

[6] https://www.digitale-edition.de.

[7] https://dig-ed-cat.acdh.oeaw.ac.at.

information to hopefully obtain a comprehensive picture of the available resources.

Against this backdrop, the mission of the Editions Registry within Text+ is to provide information about scholarly editions in a single unified system, thereby enhancing their discoverability and visibility, while also advancing access to their research data in alignment with the FAIR principles.

In doing so, the Editions Registry faces the challenge of integrating information about scholarly editions from various sources, and at varying degrees of granularity, as mentioned above. The information made available comprises a broad spectrum such as disciplinary affiliations or the languages of the underlying source materials. Importantly, the scope of the Registry extends beyond completed editions to encompass ongoing projects, provided sufficient information is available to meaningfully record them.

2.2 Scope and Application Scenarios

The Editions Registry is part of an overarching Text+ Registry, which integrates resources across various data domains and establishes connections among them. Distinguished by its cross-domain focus, the Text+ Registry serves as a pivotal research and informational tool, surpassing other systems. In essence, the Editions Registry is intended to be accessible to all interested parties, facilitating resource discovery and research endeavors. However, the specific queries users may pose and their underlying objectives vary greatly and are contingent upon individual needs and contexts. For instance, while a scholarly editor might seek exemplary practices within a specific field or discipline to inform their work, a linguist may pursue research data in a particular language, irrespective of the digital resource type (edition, collection, or lexical resource) to which it belongs. Consequently, not all information housed within the Registry will be pertinent in every instance. Conversely, given the heterogeneous nature of the editions included, it can be inferred that certain informational aspects identified for digital editions may not be applicable or verifiable for their printed counterparts within the Registry's inventory. Moreover, particularly in the case of digital editions, the accessibility of pertinent information is heavily contingent upon their life cycle.

2.3 Catering for Diversity: the Editions Registry Data Model

Developing a data model that is able to cater for the whole continuum of editions as well as ongoing projects proves challenging, since various types of relationships and granularities must be taken into account. To accommodate this vast variety of (different kinds of) editions and meet the diverse needs of the prospective users, the data model is very elaborate, but comes with a minimal model of required information at its core with further optional components to be added. One rationale behind this approach is to facilitate the inclusion of diverse catalogue data (as mentioned above) which often exhibits varying degrees of informational depth, granularity, or completeness. Additionally, adopting a minimal model aims to prevent overwhelming individuals or institutions seeking to include their editions or edition projects in the Registry but lacking extensive information or the capacity to invest significant time in manual data entry. The development of the data model involved close collaboration with representatives from various SISs, that contributed their expertise especially with regard to library catalogues and the underlying standards.[8]

Currently, the Editions Registry employs eight primary categories for cataloging, each of which is subsequently subdivided into distinct subcategories.[9] The main categories are:

- **Administrative Information**[10]: a section that provides information on the record itself (such as its provenance (e.g. AGATE, GEPRIS[11], manual), timestamp, or relations to other records for the respective edition or project
- **Output Type(s)**: a section that specifies the media form(s) in which an edition exists or will exist (e.g. printed book, electronic)
- **Basic Information**: a section that encompasses essential details such as title, year(s) of publication, existing IDs, discipline(s), status, relational metadata, and project description
- **Actors**: a section that provides information on people and institutions (funding bodies, publishers etc.) that are/were in a way involved in the making[12]
- **Information on the *edendum* or *edenda***, which are the abstract or concrete objects of editions. Among other aspects, this section informs about source medium and type of source material, period, language(s) and writing system(s), subject matters, originator(s) and/or work (if applicable).
- **Editorial Realisation**: section where – inter alia – the type of edition and its components as well as the encoding can be described

[8] Additionally, insights gleaned from the CLARIAH-DE project (2019–2021, https://www.clariah.de) have informed this endeavor. Within this context, a foundational matrix for describing and categorizing exclusively digital editions was developed, albeit in a comparatively more rudimentary form [5].

[9] https://doi.org/10.5281/zenodo.12090591

[10] Information generated largely automatically during creation.

[11] https://gepris.dfg.de/gepris/OCTOPUS?language=en.

[12] Besides an increased visibility of editions as products, the Registry also aims at elevating the visibility of those involved in the editorial process, and at raising awareness for their achievements. This is discussed in more detail in Sect. 2.4.

- **Technologies**: a section that contains information about tools or software used within the editorial process
- **Research Data Management (RDM)**: a section which informs about the RDM within the project and/or the provision of data, also with regard to the FAIR principles (e.g. use of Persistent Identifiers (PIDs), licence used etc.)
- **Additional Information** (e.g. whether a review exists)

Not all categories are (equally) relevant for all kinds of editions. Thus, references to a project's RDM and the consideration of the FAIR principles are obviously only meaningful where data is available, which is most often not the case for printed editions. Wherever possible, controlled vocabularies are used as base for the predefined selections. Additional free text fields inserted in some places, on the other hand, prevent information from being unnecessarily restricted.

2.4 Digital Editions – Registering the Makers

The creation of editions typically demands a significant investment of time and resources, often necessitating the collaboration of a diverse team, potentially spanning multiple organizations. This collective effort encompasses a range of tasks across various domains. This is especially true for digital editions, where the expertise of data modelers and research software engineers is integral, though frequently overlooked within traditional cataloging systems and beyond. One of the aims of the Registry is to make the contributions of these people visible in their entirety.

Digital editions represent a distinct genre characterized by an interdisciplinary nature. Consequently, their adequate recording in the Registry necessitates a fusion of traditional cataloging as developed by library and information science, with contemporary advancements in DH and RDM. The stakeholders engaged in digital editions encompass a spectrum of roles, ranging from *traditional* academics to digital humanists, IT professionals, data scientists, as well as project and research data managers. In many instances, individuals may embody a combination of these attributes within their educational and professional backgrounds.

In light of this context, it is essential to note that at the initial tier, the Registry for editions in Text+ primarily distinguishes between individuals and organizations contributing to digital editions. It is at the subsequent tier that the complexity described earlier is addressed. The Registry then offers an extensive array of options for assigning more specific roles to actors based on their involvement. However, it's worth mentioning that a comprehensive and precise vocabulary for characterizing these roles is not yet established. This gap also extends to TEI-XML, which is commonly regarded as the benchmark for digital editions. Consequently, various controlled vocabularies are under consideration to effectively classify actor roles:

- MARC Code List for Relators[13]
- Library of Congress relators[14]
- Dublin Core™ Metadata Initiative Terms[15]
- DataCite Metadata Schema[16]
- NFDI4Culture Ontology[17]
- CRediT (Contributor Roles Taxonomy)[18]
- FOAF Vocabulary Specification[19]
- Common European Research Information Format (CERIF)[20]

Similarly crucial is the unequivocal identification and disambiguation of the actors recorded within the Registry. This is ensured through linkage of entities to authority data such as the Gemeinsame Normdatei (GND)[21], and other external knowledge bases like ORCID[22].

The integration of these resources not only enhances the interoperability of the Registry but also fosters networking opportunities. In German-speaking regions, additional *See-Also* services, based on the GND and offered by Galleries, Libraries, Archives, Museums (GLAM) and research institutions, establish a comprehensive retrieval network. The Registry contributes to this network by referencing GND-BEACON [8] files specifically created for this purpose, and also by providing its own GND-BEACON file.

3 Technical Design

Translating the purpose and scope of a registry for Text+ resources to a technical level highlights the need for data models that can respond to the ingestion of additional catalogues, records, and future community requirements: the data model itself, the adoption of controlled vocabularies, and the integration of authority files will further shape the understanding of scholarly resources in the Text+ ecosystem and thus its manifestation as a descriptive data model.

As a directory and reference system for text and language resources, the Registry supports a variety of use cases

[13] https://www.loc.gov/marc/relators/relaterm.html.

[14] https://id.loc.gov/vocabulary/relators.html.

[15] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[16] https://support.datacite.org/docs/datacite-metadata-schema-v44-properties-overview.

[17] https://nfdi4culture.de/ontology.html.

[18] https://credit.niso.org/.

[19] http://xmlns.com/foaf/spec/.

[20] https://eurocris.org/services/main-features-cerif.

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

[22] https://orcid.org/.

for both users and connected systems. Data ingestion and modification, the creation and export of resource lists and graphs and the referencing of authority data can all be done via REST-APIs and the GUI. Amongst the many complexities, this section focuses on four key functional aspects of the Registry that have been identified as critical for early releases of the Registry: data model evolution, dynamic interface generation, catalogue consolidation and a layered resource concept.

The discussion in Sect. 2 introduced editions as one of the domains of Text+ and thus of the Registry. The examples in the remainder of this section will therefore be based on editions. The introduced complexities and challenges apply, however, to a similar extent to the domains of collections and lexical resources.

3.1 Data Model Evolution

The Registry accommodates the variability in descriptions of editions, collections, and lexical resources by supporting specific data models for each type, despite sharing common properties like title and licensing. These models are initially developed by domains within Text+, with an overarching working group ensuring early identification and management of overlaps. Given that requirements and data models will evolve, the Registry adopts a dynamic modelling principle, allowing for the flexible definition of data models. This approach enables the Registry to dynamically generate APIs for ingest and export as well as user interfaces, rather than being rigidly tied to initial data models.

3.1.1 Background

The concept of generating software or its components from formal models, such as Model-driven Architecture (MDA) originates from early Computer-Aided Software Engineering (CASE) tools in the 1980s. A study by Sebastián, Gallud and Tesoriero analyzed literature on MDA that has been published between 2008 and 2018 [6]. They concluded their study with the finding that MDA has managed to retain a prominent role in modern software development, particularly for web and GUI/API development. MDA requires strict formalization of business logic, making it suitable for complex projects needing independence from specific technologies. Its application mainly targets sectors like telecommunications, industry, and economics, where formalism is common. The humanities often deal with complex, nuanced subjects that resist simple formalization. The interpretive and subjective nature of humanities research does not easily translate into the strict formal models MDA uses.

Modern technical frameworks for form generation are model-driven and rely on the formalization of data e.g. in terms of JSON Schema[23]. There is a wide variety of form generation frameworks that can be used depending on the technology stack of an application. The Registry is based on the Spring Framework[24] at the backend, so that Spring Web MVC[25] or the more modern, reactive Spring WebFlux[26] can be used for form generation. JavaScript (JS) is used to implement the front end and there are a large number of web form generators that can be integrated based on the parent JS framework such as React JSON Schema Form[27], Angular Schema Form[28] or the Vue Form Generator[29]. From an abstract perspective, model-driven form generators apply model transformations in that data models that are specified in terms of a generic metamodelling language such as JSON schema are transformed to a representation that is defined in terms of a application-specific language – in this case HTML.

3.1.2 Resource Metamodel

For the design of the Registry we found that the formalization of the domain models – i.e. the data models for editions, collections and lexical resources – share commonalities despite differences in content, which can be manifested in editor and presentation interfaces. For this reason, the Registry has been implemented on the basis of a domain specific metamodelling language, which addresses these common features in our data models for Text+ resources. Fig. 1 shows the specification of three fields of the editions data model in terms of the Registry metamodelling language:

- `title` is a textual field, which allows exactly one entry (`mandatory==true`, `multiple==false`). Because it is also specified as a multilingual field (`multilang==true`), the exactly one restriction applies per language.
- `main_editor` is defined as field that references another resource (`entity: person`) that is defined in terms of an individual data model. The relation to this resource defines a classification vocabulary (`relation_vocabulary: _relations_person`) and in this particular case, the exact classification of the relation is predetermined in the data model (`relation_type: editor`). If the related resource is not available, a free text entry can be specified (`strict: false`) without the edition description failing validation. This is partic-

[23] https://json-schema.org/.

[24] https://spring.io/.

[25] https://docs.spring.io/spring-framework/reference/web.html.

[26] https://docs.spring.io/spring-framework/reference/web-reactive.html.

[27] https://github.com/rjsf-team/react-jsonschema-form.

[28] http://schemaform.io/.

[29] https://github.com/vue-generators/vue-form-generator.

```
entity: edition
blocks:
  - name: basic_information
    properties:
    - field: title
      type: text
      mandatory: true
      multiple: false
      multilang: true
      ...
  - name: actors
    properties:
    ...
    - field: main_editor
      entity: person
      relation_vocabulary:
          _relations_person
      relation_type: editor
      strict: false
      multiple: true
      ...
  - name: edition_objects
    properties:
    - field: source_medium
      type: vocabulary
      vocabulary:
          editions_source_medium
      multiple: true
      strict: false
      ...
```

Fig. 1 Editions datamodel

ularly useful for imports from catalogues, where authority data resolution fails to identify a match. The field allows multiple main editors to be specified (`multiple: true`).

- `source_medium` defines a field that is bound to a controlled vocabulary (`type: vocabulary`), which is in this case managed within the Registry (`vocabulary: editions_source_medium`), but could also relate to an external vocabulary. The field allows references to multiple entries in the vocabulary (`multiple: true`) and additional free text entries, if a matching vocabulary entry cannot be found (`strict: false`).

The language and its specification are currently being refined and are work-in-progress. Text+ data models for resources and vocabularies are accessible in a dedicated repository.[30]

3.2 Interface Generation

Based on specifications in terms of the metamodelling language data models and defined resource descriptions can be transformed into other representations as functionally required. Serving as a pivotal component of the Text+ infrastructure, the Registry must support bulk ingest processing, manual editing by domain experts, and the export of curated and consolidated resource descriptions via APIs.

REST-based ingest and export interfaces have been the main focus of API development for early versions of the Registry, as they are used for client/server communication of the user interface of the Registry itself. REST and JSON are now frequently used in DH services and complement XML as widely adopted protocols. Both the ingest and export interfaces of the Registry deserialize and serialize data that is defined in terms of specified data models. For the adaption to heterogeneous formats as e.g. data provided by the aforementioned catalogues, the Registry builds on two infrastructure services of Text+, the Data Modeling Environment (DME) for data modelling and mapping [7] and the Transformation Service (TS) for the actual transformation on the basis of the models and mappings [2]. In future releases, the Registry – possibly with the help of the DME and TS – will provide data in terms of other formats, such as the de-facto-standard for data sharing Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[31], and Resource Description Framework (RDF)[32] to be able to export resource descriptions within their context and relations.

With regard to the user interface, Fig. 2 shows the representation of the fields that have been specified in Fig. 1. Interface rendering is executed on the backend on the basis of the metamodelling language, the specified data model and the resource description. As the *Title* field allows the formulation of a title per language, controls for the addition and removal of input elements are provided along with a typeahead-based selector for the language. The relation of a *Main editor* is created by the same type of selector that searches matching resource descriptions. And the *Source medium* allows the selection of entries within the configured controlled vocabulary with the option to provide free-text.

The metamodelling language is used to declaratively formulate further characteristics of a data model. One particular example is the extensive validation framework that has been implemented for the Registry. The combination of possible rules, such as the title's `mandatory==true, multiple==false, multilang==true` requires a sophisticated and conditional combination of validation

[30] https://gitlab.com/minfba/resinfra/textplus-registry-models.

[31] https://www.openarchives.org/pmh/.

[32] https://www.w3.org/RDF/.

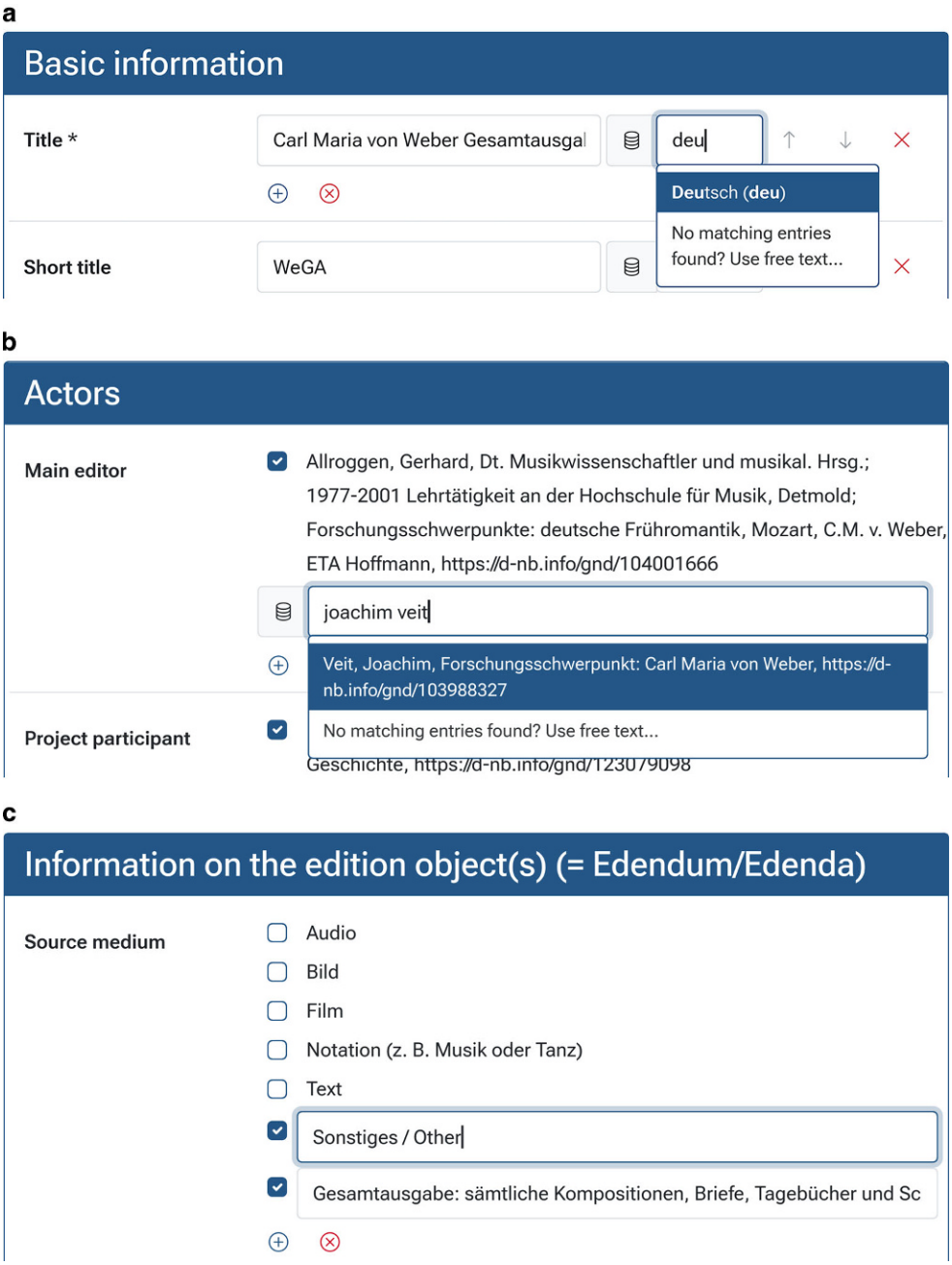

Fig. 2 Elements in the Registry editor. **a** Multilingual text field `title`, **b** Reference to other entity `main_editor`, **c** Controlled, non-strict vocabulary `source_medium`

statements. In addition, validity of a resource description is not a binary statement for the Registry as content can be schematically valid, but semantically invalid. An example is a valid URL that leads to a 404 NOT FOUND server response. The Registry should allow users to save the description, but notify them of the flaw.

Another feature that is based on the metamodelling language is localization of the user interface. Data models and field names describe a naming convention that is used for the translation of labels. Domain experts can formulate translations e.g. for a code `edition.title` and register them in an instance in Weblate[33], a libre continuous localization software. Translations of data models and fields can be recorded without tampering with the source code or data model specifications.

3.3 Consolidation of Catalogues

A key objective of the Registry is to consolidate descriptions of text and language resources. It will build on the data provided by existing catalogues and repositories and facil-

[33] https://weblate.org.

itate the transparent combination of this data. In addition to the re-use of already existing information, the consolidated resource descriptions can be refined by experts in the respective domains, and the Registry will be a central point-of-access to Text+ resources. For initial releases of the Text+ Registry the correlation of descriptions of the same edition in multiple catalogues is treated as a manual task to be performed by domain experts. The evaluation and robust application of duplicate detection algorithms to support this correlation is an area for future work.

3.3.1 Architecture

The design of the Text+ Registry builds on the idea that descriptions are already available for some textual resources. Data repositories often provide metadata on contained resources. Accessible details range from very basic metadata e.g. by using the `set` construct of the OAI-PMH specification to explicit metadata records on resources. As one of many examples in Text+, the Language Archive Cologne provides information on contained resources in terms of the Basic Language Archive Metadata Bundle[34], a Component Metadata Infrastructure (CMDI)[35] profile, which allows the formulation of comparably extensive resources descriptions. In addition to metadata that can be obtained directly from repositories, catalogues collect and provide access to comprehensive datasets, often with a particular community or disciplinary focus. Catalogue data can be used to enrich existing metadata provided by the repository, or to replace missing repository metadata for non-digital editions or other resources for which the respective repositories do not provide metadata.

3.3.2 Example

An example of how catalogue data can be used to enrich the already rich metadata provided by the original source is WeGA. With respect to the availability of project and edition metadata, WeGA can be considered a showcase project in that a large descriptive set of metadata in the form of TEI-XML is provided on the project website. Despite the richness of the provided metadata, the inclusion of catalogue data leads to a more extensive description of the edition.

To illustrate the benefit of combining collection-level metadata from multiple sources and the resulting requirements for the Registry, we consider data from three catalogues with corresponding entries for WeGA. Please note that despite their value to their respective communities, the three catalogues are only presented as examples of the context of editions, and there are many other sources of resource descriptions that may be added to the Registry in the future:

- **AGATE** understands itself as gateway for the Humanities and Social Sciences and contains information on research projects that have received funding by the German Academies Programme since 1979 [10].
- The **CDE** collects digital editions and texts. It is tailored towards cataloging digital editions and details digital projects with titles, descriptions, URLs, and possibly technical details like digital features, scholarly aspects, and infrastructure used [1].
- **DSE** contains detailed textual data, metadata, and annotations for DH projects, emphasizing scholarly texts, manuscripts, and historical documents [4].

Table 1 Comparison of machine-readable metadata provided on WeGA

	AGATE	CDE	DSE	WeGA
Title	+	+	++	++
Abbreviation, Acronym	+	–	(+)	(+)
Description	++	–	+	++
Actors	++	+	(+)	++
Institutions	+	++	(+)	(+)
Date (Range) of Publication	+	–	+	+
Reviews	–	–	+	–
Catalogue ID s	–	++	+	–
URLs	++	+	+	+
API available	–	+	(+)	(+)
Contact Info	–	–	+	+

Table 1[36] compares the metadata available in the three catalogues and the project website. Despite the comprehensive nature of the project description, the catalogues provide additional information. For example, AGATE contains a wealth of curated annotations detailing topics, disciplines and DH methods, and it details personal relationships beyond project leadership. DSE refers to the recognition in *A review journal for digital editions and resources (RIDE)*[37] and provides technical details for accessing the edition. In addition, even if no additional explicit metadata were provided by the catalogues, their contextual setting allows conclusions to be drawn about the properties of the edition, since all the projects referenced in AGATE have been funded by the German Academies Programme, and

[34] https://api.ka3.uni-koeln.de/oai/lac?verb=ListRecords&metadataPrefix=cmd.

[35] https://www.clarin.eu/content/cmdi-component-metadata-infrastructure.

[36] Overview of machine-readable catalogue information on the WeGA edition (partial) and the WeGA project description. Due to the different objectives of the catalogues, the metadata offered varies in scope and detail – denoted from "++" (in detail or in full) to "+" (in part or in excerpts). Sometimes aspects are only indirectly covered "(+)" or not available "–".

[37] https://ride.i-d-e.de/issues/issue-12/wega/.

the editions in the CDE and DSE are digital editions. As a result, we concluded for the design of the Registry that the integration of multiple sources leads to richer resource descriptions in the Registry – as opposed to just one catalogue or the project source itself – potentially reducing the effort required to manually enrich descriptions to meet the needs of the domain data models.

3.4 Layered Resources

To support the consolidation of resource descriptions from multiple sources, the Registry implements descriptions as *layered resources*. Essentially, this concept involves stacking available descriptions atop one another. Each layer remains transparent regarding properties it does not address but supersedes properties it does address. Fig. 3 illustrates the concept of layered resources based on the three catalogues described above: from the data imported from the DSE, only the review is retained as titles and persons are overridden; of the data imported from the CDE, only the technical metadata is seen in the consolidated resource description, and so on. After consolidation of imported resource metadata, manual annotations are used to fill details that have not been imported in any of the subordinate layers.

The layered approach implements features that we believe are essential to the design of the Registry and its acceptance by the research communities: *robustness of updates* and *provenance*. Since the integration of multiple catalogues into a single record inevitably results in a loss of information and provenance, individual catalogue data must be preserved in parallel – an observation that has also been made by Sahle in his work on the Codices Electronici Ecclesiae Coloniensis (CEEC) project [3].

As the Registry imports resource descriptions into distinct layers, these imported layers remain shielded from alterations caused by imports from alternative sources or manual curation. Each layer is inherently associated with a specific import source, thereby restricting content modifications exclusively to updates originating from the same source. Consequently, this mechanism fosters a heightened level of transparency concerning the provenance of specific data. When generating and showcasing a consolidated resource description for the API or user interface, references to the source of individual properties of the resource can be provided with exceptional granularity.

Currently, domain experts possess the capability to arrange the layers' order within the stack according to their preference, populate data within the manual annotation layer, and thus override, augment, or conceal imported data. However, there may be a necessity for more sophisticated and nuanced functionalities, such as selectively highlighting specific properties from an otherwise ordered lower layer. These enhancements could potentially be incorporated into the Registry in future iterations.

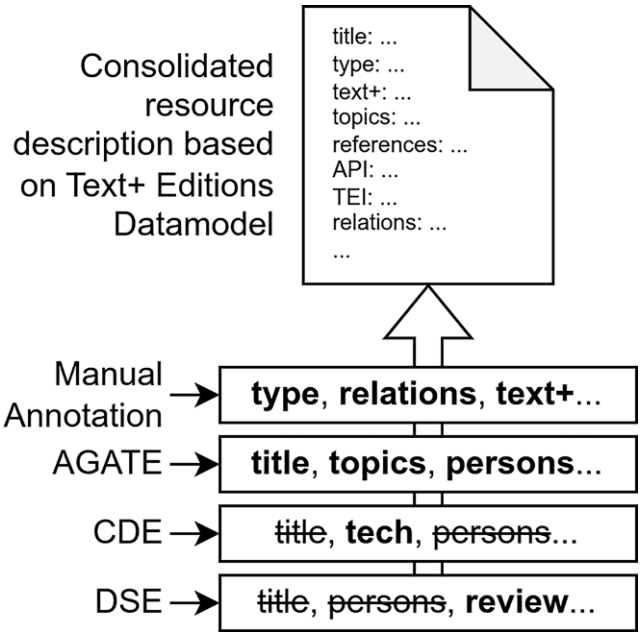

Fig. 3 Description layering for editions

4 Conclusion

Discovering scholarly resources remains a challenging task for scholars in the DH. Resource descriptions are scattered across multiple catalogues – each with a different disciplinary focus and data model. Descriptions are often not classified in ways that allows immediate filtering and searching, and lack integration with knowledge bases such as authority files.

The development of the Text+ Registry is a significant step towards addressing the challenges of making scholarly resources more FAIR. This paper has detailed the technical design of the Text+ Registry a central platform intended to facilitate the discovery and accessibility of digital resources across different academic disciplines. We have presented an approach that focuses on the application of a metamodelling strategy to facilitate the formulation and evolution of domain data models, the automatic transformation of data models into APIs and GUI components, and the provenance-focused consolidation of existing catalogues based on the conceptual foundation of layered resources. We consider these efforts as critical for overcoming the inherent challenges in ensuring that digital resources are easily discoverable and usable by researchers across fields.

The technical design and the architectural choices that underpin the Text+ Registry, represent practical steps to-

wards implementing and improving the functionality and effectiveness of the Registry.

However, the paper also recognizes the ongoing nature of this work. Future directions for the Text+ Registry, as outlined in the paper, include not only technical improvements but also an increased engagement with initiatives such as the SISs to ensure that architectural and technical choices match the needs and objectives of scholarly communities.

Acknowledgements This document was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the DFG – project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

Author contribution Conceptualization: all authors; Writing – original draft preparation: all authors (Sect. 2 mainly by CK, HL and DS; Sect. 3mainly by TG); Writing – review and editing: all authors; Software: TG

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Franzini G, Mahony S, Terras M (2016) A catalogue of digital editions. Open Book Publishers https://doi.org/10.11647/OBP.0095
2. Jegan R, Fruth L, Gradl T et al (2023) Integrating access to authority data for improved interoperability of research data in the digital humanities https://doi.org/10.18420/BTW2023-54
3. Sahle P (2001) Erfahrungen mit internationalen standards bei der handschriftenbeschreibung. Codices electronici ecclesiae coloniensis Eine mittelalterliche kathedralbibliothek in digitaler form, pp 77–95
4. Sahle P (2016) What is a scholarly digital edition? Open Book Publishers, pp 19–40 https://doi.org/10.11647/obp.0095.02
5. Schulz D, Fisseni B, Sendler S (2021) Edma – eine matrix zur erfassung und kategorisierung digitaler editionen. Tech Rep. https://doi.org/10.14618/IDS-PUB-10501
6. Sebastián G, Gallud JA, Tesoriero R (2020) Code generation using model driven architecture: a systematic mapping study. J Comput Lang 100(56):935. https://doi.org/10.1016/j.cola.2019.100935
7. Steyer T, Gradl T (2019) A research-oriented and case-based data federation for the humanities https://doi.org/10.5281/ZENODO.2536107
8. Voss J, Schindler M (2017) Beacon link dump format. Tech. Rep., Network working group. https://gbv.github.io/beaconspec/beacon.html. Accessed February 19, 2024
9. Wilkinson MD, Dumontier M, Aalbersberg IJ et al (2016) The fair guiding principles for scientific data management and stewardship. Sci Data. https://doi.org/10.1038/sdata.2016.18
10. Wuttke U, Ott C, Adrian D et al (2017) Agate: concept for a european academies internet gateway for the humanities and social sciences https://doi.org/10.5281/ZENODO.806179

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.