Cataloguing editions and other resources in one unified system – The case of the Text+ Registry

By Leon Fruth, Nils Geißler, Tobias Gradl, and Daniela Schulz

1. Introduction and Bigger Picture

Text+ is a consortium of the German National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur, NFDI). With its 2016 recommendation, the German Council for Information Infrastructures (RfII) initiated the establishment of the NFDI. To solve the problem that data is mostly decentralized, project-related and/or only available for a limited time, the NFDI wants to create a permanent digital knowledge repository for Germany. Relevant data from all disciplines shall be made available in the long term according to the FAIR principles *findable*, *accessible*, *interoperable* and *reusable* (cf. Wilkinson et al., 2016).

To cover the national science system in all its breadth, a consortium structure was chosen. Consortia, in this sense, are associations of different institutions within a wider field of research that work together on an interdisciplinary basis in order to achieve a common set of objectives. A total of 26 consortia is currently funded as well as a cross-sectional consortium called Base4NFDI, responsible for common services relevant to all consortia. Four consortia address the area of humanities: NFDI4Culture for research data on material and immaterial cultural heritage, NFDI4Memory, representing the field of history and related disciplines, NFDI4Objects dedicated to the interdisciplinary study of material remains, and Text+.

Regarding its scope, Text+ is focused on the sustainable preparation, provision and preservation of text- and language-based research data and aims at constructing a flexible and scalable research data infrastructure able to cater to the unique requirements of academic disciplines that work with those data. It is organized along the three data domains collections, lexical resources, and editions. Its objective is to address the respective scientific communities and to cooperate with them on the further development of standards and best practices to make data available in accordance with the aforementioned FAIR principles. To do so, Text+ provides a whole range of services, data and consulting not only for institutions and associations, but also projects and individual researchers.

A core element of the Text+ infrastructure is the Text+ Registry, which is intended to be a unified system for describing and cataloguing different types of resources from the three domains. While existing databases and catalogues often come with a very specific focus (discipline, type of resource, etc.) and lack features for integration with external sources, the Text+ Registry goes beyond this. It is able to work with different data models using a metamodeling approach, and allows the import, curation and provision of resource descriptions. Other search and retrieval tools such as the CLARIN Federated Content Search (FCS) and the DARIAH-DE Generic Search (GS) are integrated. APIs for data ingest from existing catalogues (like Greta Franzini's Catalogue of Digital Editions (CDE) and Patrick Sahle's Catalog of Digital

in: Zhou, Lorraine (Ed.). Proceedings of the Digital Humanities Congress 2024. Sheffield: The Digital Humanities Institute, 2025. Available online at: https://www.dhi.ac.uk/books/dhc2024

Scholarly Editions (CDSE), library or union catalogues, research information systems of funding bodies such as AGATE[1]) and data provision are also supplied. The Text+ Registry should not only increase the visibility of scientific resources, but also overcome the traditional barriers that hinder their discoverability and accessibility.

2. Status quo: Desiderata and Challenges

While editions constitute the foundation for scientific research and discourse, their discoverability can be a difficult task at best. There are a number of reasons for this: scholarly editions usually result from third-party funded projects, and not only the nature of the editions, the parties involved, but also the funding requirements and dissemination formats vary widely. The research information systems of funding bodies such as the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) or the Academies' union contain some information on these projects, but usually lack integration with external knowledge bases. Printed editions in Germany are generally recorded in library catalogues, but tracking them down is not trivial as there is no commonly used subject term for them. Digital editions are for the most part not recorded in library catalogues at all. The aforementioned inventories of digital editions that are curated by individuals or small groups, come with specific scopes and hence provide different (levels of) information. As a result, users will find bits and pieces of information in different places, but will have to manually query many systems.

It can be summarised that the landscape of editions as a whole is diverse and quite complex, making it hard to obtain at least comprehensive coverage. Considering the number of disciplines active in the field of scholarly editing, the numerous approaches, practices and traditions that have evolved over the centuries, or the deliberate rejection of these, this is only logical. The emergence of new technical possibilities has further fuelled this trend. According to Patrick Sahle, each editorial subject has its own requirements and thus leads to more or less individual solutions (Sahle, 2013, p. 12). Developing a classification system able to cater for the whole continuum of existing editions as well as ongoing projects, and also for working with different information sources and levels of informative content, thus proved challenging. Another difficulty lies in the sheer number of relevant editions that do exist.

3. Approach: Towards an Editions Registry

The Text+ Registry aims to provide structured, comprehensive, and context-aware descriptions of text- and language-based resources from the three data domains collections, lexical resources, and editions. Within this broader framework, the domain of editions presents a particularly heterogeneous and technically demanding field. They differ widely in terms of their media formats, editorial scope, historical source materials, and disciplinary conventions. To meet these challenges, a dedicated **editions data model** was developed in close collaboration with experts and stakeholders from research, infrastructure, and library communities. This model provides the conceptual foundation for consistent and meaningful cataloguing across a wide variety of edition types. In parallel, practical procedures for **cataloguing and data**

integration were established to support both automated ingestion and manual enrichment. These procedures allow the registry to consolidate metadata from existing catalogues, while also encouraging direct contributions from researchers and institutions. Together, the data model and the cataloguing strategy constitute a modular and extensible framework for describing editions in a way that supports discoverability, reuse, and long-term curation—aligned with the goals and principles of the Text+ infrastructure.

3.2 Editions data model: Scope, Development and Procedures

Regarding its **scope**, the Editions Registry is per se not limited, except the basic prerequisite that the resource to be described is an edition, including both ongoing projects as well as completed editions of any media form. However, sufficient information must be available to create an entry that at least covers the few mandatory fields (i.e., minimal entry) of the data model as outlined below. While ensuring the quality of the metadata provided is of significant importance, and hence a lot of effort is put also into the manual curation of the records, no evaluation of the editorial quality is carried out.

In principle, the registry should be available to all interested parties as a **research tool**. It is intended to satisfy a wide range of scholarly needs, such as identifying best practice examples for orientation, finding relevant resources for teaching purposes or data for textual analysis, or increasing the visibility of one's own work. What users search to find and for what purpose is thus highly context-dependent and not all the information held in the registry will be relevant in any case. Conversely, due to the heterogeneity that is prevalent, it is likely that not all information components that are of relevance to digital editions will also be transferable or verifiable for print editions. Furthermore, especially in case of digital editions, the availability of the respective information components is also strongly dependent upon their life cycle.

The **data model** was developed in several stages. First, the working group reviewed relevant previous work and compared and consolidated the respective categorisations that were applied there.[2] On this basis, a preliminary data model was designed and tested for its suitability to describe different editions using representative data sets. An initial contribution of data to the Text+ Registry came from the institutions that are part of Text+ and that built the so-called **portfolio**.[3] This portfolio did not only serve to shape the data model as test data, but it was also a way to prompt further data in a higher level of detail from experts who actually worked within the corresponding projects. In addition to the portfolio, the registry collects metadata on editions regardless of institutional affiliation. While no distinction is made between editions provided by Text+ partners and those from other institutions, institutional affiliation is explicitly recorded as part of the metadata. The model was repeatedly presented during its development phase and introduced to experts outside Text+ for discussion. Representatives of the various specialised information services (SIS, in German: Fachinformationsdienste, FID[4]) in particular contributed their expertise with regard to library catalogues, underlying standards and interfaces and thus helped to fine-tune the editions **data model**.

For its development, various types of data, relationships and granularities had to be taken into account. This, together with the aims to accommodate the vast variety of editions, and also to meet the diverse needs of prospective users, resulted in a very elaborate model (cf. Measure 1 der Task Area Editions, 2024).[5] The main categories employed for cataloguing editions, which consist of further sub-categories, are as follows:

- **Administrative information:** this block contains information about the data record itself (e.g., its provenance, information about the time of its creation or relationships to other data records)
- **Output Type(s):** a section that specifies the media form(s) in which an edition exists or will exist (e.g., printed book, electronic)
- **Actors:** a section that provides information on people and institutions (funding bodies, publishers etc.) that are/were in a way involved in the making (making use of controlled vocabularies and authority files, such as the Integrated Authority File (Gemeinsame Normdatei, GND) for the disambiguation)
- **Editorial Realisation:** section where – inter alia – the type of edition and its components as well as the encoding can be described
- **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).
- **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, licences used, etc.)

The data model (see Figure 1) is the foundation of the HTML forms and layouts you see on the registry website (see Figure 2).

Hierachy	Level	Filed label	Function	Selection	Help	DataCite reference	Comment
0	0	Administrative Information					This section contains information about the record itself (e.g. its provenance, information about the time it was created or relationships to other records).
0.0	1	resource type	1	lexical resource, collection, edition, service	Please indicate which resource type you are dealing with.	10. ResourceType — DataCite Metadata Schema 4.5 documentation (datacite-metadata-schema.readthedocs.io)	
0.1	1	Text+-ID	1				
0.2	1	provenance	1	manual, AGATE, GEPRIS, CSE, CDSE, curated	The data source or provenance from which the entry originated can be specified here. In the case of "manual entry", this entry itself is the provenance.		
0.3	1	timestamp	1				
0.4	1	versions	0..n		The versioning of the registry entry is displayed here. Previous/subsequent edition versions of the edition itself can be specified under "Relations".		
0.4.1	2	versions	0..n				
0.4.2	2	type of relation	1				
0.5	1	part of the Text+ portfolio	1	[x]			
1	0	output type					This section contains information on the output type(s) in which the edition is or will be available.
1.1	1	type	1	book version (print/ebook), digital edition (portal, webpresentation), hybrid, research data	Please select the output form of the edition. If there are different forms of edition, you can enter these in separate entries and then link these entries via "Relations".		
1.2	1	citation	0..n		You can enter a recommended citation here.		
1.3	1	language of output type	0..n		Enter the language(s) used in the output form here. The languages that occur within the edition itself are recorded under "Information on the edition object(s)".		
1.4	1	licences	0..n			16. Rights — DataCite Metadata Schema 4.5 documentation (datacite-metadata-schema.readthedocs.io)	
1.4.1	2	information on accessibility	0..n	free, free for academic purposes, restricted, not accessible	Specify whether the resource is free or only accessible with restrictions.		
1.4.2	2	licence	0..n	CC, free text	Select a licence from the list or use the blank text field.		
2	0	basic information					This section contains basic information about the edition or project, such as title, publication year(s), existing IDs, discipline(s), status, relational metadata and a short project description.

Figure 1: Excerpt from the data model

Figure 2: Overview of the properties as shown on the Text+ Registry website for each edition record

As shown in Figure 3, although not all categories are (equally) relevant for all kinds of editions, (nearly) complete entries can obviously become very long and detailed due to the aforementioned elaborateness of the data model.

Figure 3: Record of "The Edition of the Frankish Capitularies" as an example for a comprehensive and curated entry in the registry (rotated)

It therefore comes with a **minimal set of required information** at its core, to which other optional components can be added. One consideration behind this concept is to facilitate the inclusion of diverse catalogue data that come with varying degrees of informational depth. In addition, the adoption of a minimal model is intended to avoid overburdening individuals or institutions who wish to add their editions or edition projects to the registry but do not have extensive information or the capacity to invest substantial time in manual data entry. As a special concession for people who want to enter a minimal entry for an edition without much effort on their part, a postcard (see Figure 4 below) has been developed on which the relevant information can be submitted and that simultaneously serves as an advertising or information flyer for the Editions Registry to engage with the community, e.g. at conferences.

Figure 4: Print template of the postcard

An entry in the registry can therefore be of basic extent at first and expanded later. This means the initial entry can be done by any person with access to the limited information needed for the minimal entry. An **optional enrichment** in the course of a curation procedure can then be carried out later on also by experts closer to the project or discipline, Text+ staff members, or for example within teaching or workshop contexts. Although this iterative approach in itself is trivial, it demonstrates the values of openness, active cooperation with the scientific community for quality assurance and enhancement, and thus sustainability due to enabling a division of labour. Of course, this also requires appropriate versioning and redaction of the entries. As an attempt for **event-driven curation**, so-called editathons have been introduced and tested to bring together experts to enrich existing entries. The concept of editathons (or edit-a-thons) is based on hackathons and is also successfully used e.g. by the Wikimedia Foundation to enrich their data.

Wherever possible, the registry relies on the use of **controlled vocabularies** and **authority files.** Information on persons, institutions and works includes - as far as available - the corresponding GND numbers. For project participants who do not have a GND number, the use of ORCID is planned as an alternative. The Contributor Role Taxonomy (CRediT) and the MARC Code List for Relators (at least in parts) are currently tested to attribute roles and tasks. The former taxonomy in particular is currently becoming increasingly established for standardised information on the participation of individuals,[6] but both do not fully cover the area of editions or are too broad overall. Assignment to subject disciplines is based on the so-called Basisklassifikation (BK), a classification established in the library context and the DFG Classification of Scientific Disciplines (DFG Fachsystematik).[7] DataCite is used as a common basis across the domain-specific data models to enable cross-domain searches and queries.

3.2 Cataloguing procedure(s)

For the inclusion of data, a fundamental distinction must be made between automated procedures and manual input or editing (i.e. curation and enhancement) of entries. A complementary approach seems sensible. The **procedure for the inclusion** of the metadata of editions consists of the following modules, which do not necessarily happen one after the other but often in parallel: preparation of as complete as possible entries for Text+ related editions (i.e. portfolio data), systematic inclusion of further editions and edition projects in order to provide a balanced set of data from various disciplines (semi-automated/manual), import of a variety of catalogue data, enhancing, improving and correcting of (automated) entries, importing/harvesting information from further relevant lists and databases, enabling entries to be made by external scholars (also as part of organised events and workshops already mentioned).

In the **landscape of digital editions** you will eventually come across two catalogues, which are usually just referred to as the "Sahle catalogue" - "the first ever catalogue of digital editions"[8], dating back to the late 1990s[9] - (introduced as CDSE) and the "Franzini catalogue" (introduced

as CDE), each of which attempts to provide an exhaustive overview of said landscape. While the CDSE has a minimal data model, the CDE is more verbose and even states whether an edition is also present in the CDSE. Non-digital editions or editions that have not been published in digital form, are by intention not registered in either catalogue due to their focus domain. As another example, the database of AGATE (cf. Wuttke et al., 2017) contains information on research projects that have received funding by the German Academies Programme since 1979. In this way, each catalogue has its own perspective on editions, which is reflected in its metadata. A detailed analysis of these catalogue entries and the associated machine-readable descriptions provided on the project's website can be found in Gradl et al. (2024, pp. 158f.).

- CDSE entries correspond specifically to digital presentations of scholarly editions, using TEI-XML to describe essential bibliographic details, publication dates, and classifications such as material type, language, era, and subject area. CDSE also enriches entries through cross-references to academic reviews, for example, those published in *A Review Journal for Scholarly Digital Editions and Resources (RIDE)*.
- CDE similarly captures digital scholarly editions but extends the metadata focus to technical and digital presentation aspects. A machine-readable form of CDE entries is available as a CSV document. While overlapping with CDSE in terms of fundamental descriptive elements, it provides additional structured data about editors, institutional contexts, accessibility of images, availability of APIs, and downloadable TEI-XML files.
- AGATE, by contrast, adopts a distinctly project-oriented perspective. It centres metadata descriptions around projects funded by the German Academies Programme, focusing primarily on the editorial and research context rather than the detailed properties of individual publications. Consequently, AGATE entries emphasise funding information, project statuses, research objects, methodologies, and team compositions.

Combining these perspectives, each with distinctive scopes and metadata emphases becomes a desideratum.

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

Figure 5: Comparison of the varying metadata (Gradl et al., 2024)

4. The Text+ Registry as Catalyst for FAIR data

Resource registries play a crucial role in enhancing the FAIRness of resource descriptions, thereby improving the usability and impact of the resources themselves. In the context of the Digital Humanities, they collect, organise and provide access to metadata about digital resources relevant to humanities research and education. Such resources can include datasets, digital tools, projects, publications, educational courses and more. For the case of the Text+ initiative, resources that are relevant to text- and language-based research are categorised along the three Text+ data domains. From a pragmatic perspective, the Text+ Registry functions as a component that improves the alignment of resources with the FAIR principles, enhancing their findability, accessibility, interoperability, and reusability.

Acting as a centralised index, it facilitates **findability** of resources that are spread across different systems and organisations. Comprehensive data models and advanced search functionalities enhance resource discovery and location, particularly benefiting the decentralised Text+ data landscape. Here, resources not only span multiple repositories but are also relevant across various academic disciplines beyond their original context. Since registry entries can combine multiple source descriptions of existing catalogues, unique and persistent identifiers (such as DOIs or URIs) are aggregated and available for resource identification and discoverability. Activities of curation such as the classification of data along pre-existing or developed vocabularies and their contextualisation on the basis of authority files further increase findability of resources.

Resource descriptions and access methods from existing catalogues and data centers inherently differ due to varying institutional practices, metadata standards, and technical implementations. The Text+ Registry is built on the foundation of sophisticated infrastructure components to be able to consume available data independent of both the technical mechanisms used for data access (e.g., OAI-PMH, Git repositories, web scraping) and the metadata schemas and formats in which data are provided. Providing data through open access mechanisms such as standard APIs for harvesting and querying data, the **accessibility** of resources is improved. The registry ensures that metadata remains accessible, clearly specifying access conditions for the resources themselves, and providing reliable mechanisms to retrieve metadata even when direct access to the resource may be restricted.

Text+ data domains define tailored data models that capture the level of detail necessary to support Text+'s role as a specialised infrastructure provider for text and language-based research. These data models balance specificity—ensuring sufficient context and discoverability—against overly granular descriptions that apply only to narrow resource subsets. The registry promotes **interoperability** through the extensive use of standardised vocabularies within data model specifications and the publication of data in line with standardised metadata schemas. The same infrastructure components that are being used to access heterogeneous data from external catalogues are in place to seamlessly transform data as represented in the registry data models to other formats. Currently, all resource descriptions in the registry can be accessed in conformance to the Dublin Core (DC Simple) and DataCite schemas. Additionally, the registry prepares resource descriptions for integration with semantic infrastructures and knowledge graphs. This includes explicit contextualisation through structured entities,

relationships, and mappings to authoritative identifiers, enabling advanced semantic querying and integration into broader Linked Data environments.

Reusability is actively supported through detailed, provenance-aware metadata that transparently documents the origin, context, licensing, and usage conditions of resources, facilitating confident reuse in diverse academic and interdisciplinary contexts. By integrating descriptive layers and encouraging expert curation, the registry ensures high-quality metadata that can reliably be reused in various contexts beyond their original scope, thereby supporting long-term data sustainability. User interfaces as well as the APIs of the registry are designed to broadly satisfy requirements for accessing, harvesting and querying the resources.

Collectively, the Text+ Registry has been designed to systematically align resource descriptions and functionality of the registry with the FAIR principles, effectively enhancing the quality, sustainability, and impact of resources and emphasising the practical application and sustainability of digital resources across academic disciplines.

4.1 Existing catalogues as a base layer

Traditionally, resource registries can be classified into two main categories: those that reference external descriptions from external sources (such as WorldCat) and those that maintain independent internal records (e.g., CDSE, CDE and AGATE). Each approach offers distinct advantages: referencing external descriptions promotes consistency, reusability, and efficient updating, while maintaining internal records enables detailed, domain-specific customisation and expert curation.

A fundamental and distinguishing characteristic of the Text+ Registry is its design as a **hybrid catalogue**, combining these two approaches. Similar to conventional aggregators, the registry ingests resource descriptions from external sources and integrates them into a unified index. At the same time, it functions as an authoritative catalogue in its own right, permitting manual description and curation of resources not captured by external sources – either because they are not known or because they lie outside the defined domains of existing catalogues, such as particular disciplinary, language or national contexts.

The hybrid mode enables the Text+ Registry to consolidate diverse perspectives from multiple external sources, such as the edition catalogues outlined above. Consolidating resource metadata from these complementary catalogues achieves several significant benefits. Firstly, it enables the reuse of existing descriptions, reducing redundant editorial effort and minimising maintenance costs since updates from source catalogues propagate automatically into the Text+ Registry. Secondly, integrating multiple perspectives results in richer, more nuanced, and comprehensive metadata descriptions. Finally, this approach enhances long-term sustainability by ensuring that responsibility for metadata maintenance is distributed among contributing external catalogues, positioning the Text+ Registry as a central interface for expert review, enrichment, and contextualisation within text- and language-based research.

As a hybrid catalogue, the registry also provides a practical solution to existing gaps in the broader landscape of resource descriptions. It explicitly supports horizontal enrichment—expanding the registry's scope by adding resources from additional external catalogues or through manual entry—and vertical enrichment, deepening existing descriptions by aggregating complementary information. Since resource sets in external catalogues are rarely mutually exclusive, inherent overlaps frequently occur. Variations in descriptive elements result from differing cataloguing perspectives and metadata schemas, each highlighting distinct but complementary characteristics. Some schemas prioritise broad, generic, and universally applicable metadata elements, while others offer highly structured and domain-specific information.

For instance, the *Weber Complete Edition* is documented across the three catalogues mentioned above, each highlighting unique facets. Thus, they complement each other horizontally—by including entries that are absent in the other sources—and vertically—by adding depth to resource descriptions through their unique, specialised perspectives.

4.2 Provenance-aware Metadata Composition

A primary design principle of the Text+ Registry is the reuse and consolidation of existing resource descriptions where feasible. By integrating metadata from authoritative external sources, the registry reduces redundancy, leverages pre-existing curation and automatically propagates updates from these sources. This approach distributes the curation workload efficiently across connected catalogues, ensuring sustainability and scalability of metadata maintenance. To support effective metadata integration, the registry implements a sophisticated metadata layering mechanism that explicitly documents metadata origin, provenance and subsequent modifications. This provenance-aware approach enhances transparency and facilitates incremental enrichment and refinement of metadata without compromising the integrity of the original metadata sources.

Figure 6 illustrates the role of the Text+ Registry as a central hub interconnecting sources of resource descriptions, domain experts, authority data providers and data-consuming agents or aggregators. The availability and selection of relevant ingest sources vary with the resource types managed within the Text+ Registry. While edition descriptions are broadly ingested from existing catalogues as outlined above, collections and lexical resources are primarily described using data sourced directly from connected Text+ Data Centres. Through tailored data models for each resource type, the registry flexibly specifies the vertical depth of metadata descriptions. Moreover, the fine-grained definition of ingest sources helps clearly delineate the breadth and thus the horizontal scope of registry coverage.

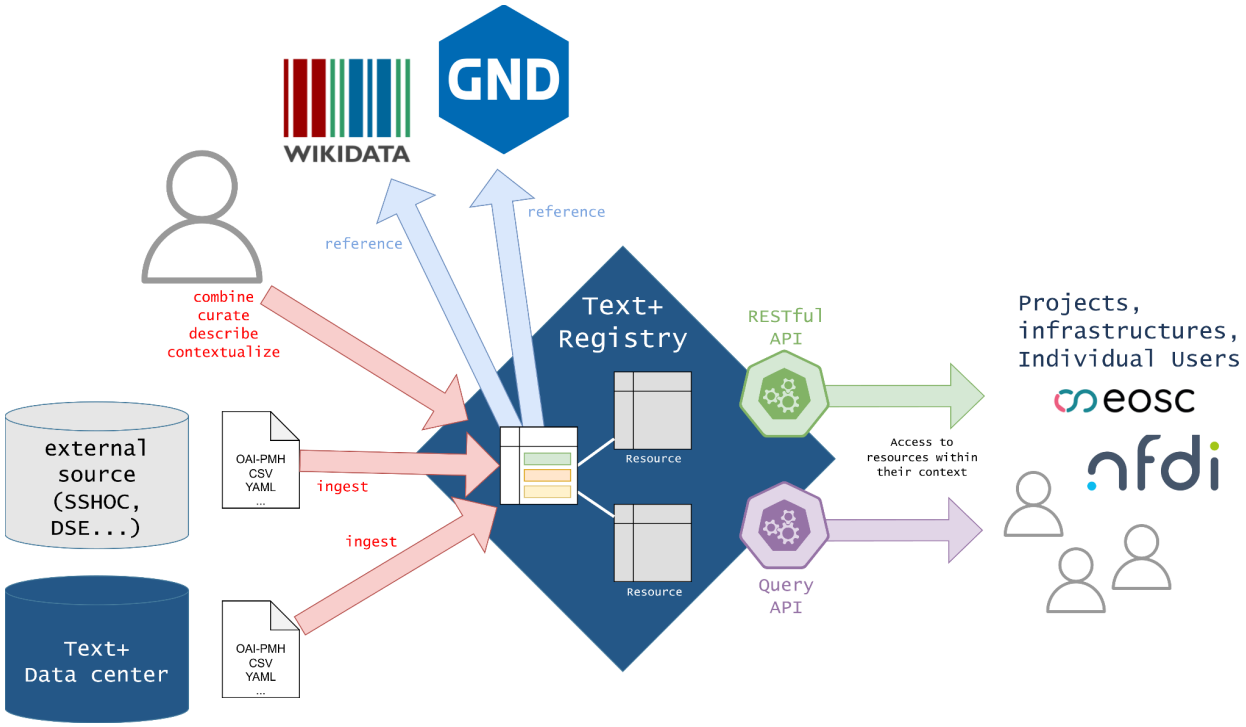

Figure 6: The Text+ Registry as an enrichment and accessibility hub

Ingested resource descriptions are managed as isolated description layers, each bound to its originating source and provenance context. Modifications within these layers can only be introduced via updates from the original data sources. Domain experts utilise these layers to identify, specify, and manage overlaps among resources, effectively curating comprehensive metadata descriptions. Layers are then dynamically stacked according to customisable priorities, determining field precedence from each source. Figure 7 exemplifies this layering principle, illustrating prioritisation among metadata sources for digital editions. In this example, metadata ingested from AGATE is prioritised over information from the CDE, which in turn is prioritised over the CDSE. For instance, the consolidated description of an edition would represent the title as imported from AGATE. However, if certain metadata fields (such as review) are unavailable in the higher-priority layers, corresponding fields from lower-priority sources (here e.g., CDSE) are included in the consolidated metadata representation.

Two critical aspects of the registry layering approach should be emphasised:

- Even if specific fields from lower-priority layers are overshadowed in the primary consolidated representation, all ingested metadata, along with explicit provenance annotations, remain transparently accessible as secondary components of the metadata record.
- The registry defines a default source prioritisation per resource type. However, this prioritisation can be explicitly customised by domain experts on a per-record and even per-field basis through an additional curation layer. Thus, curators can selectively reorder source priorities for metadata fields or manually formulate entirely new metadata entries if existing descriptions do not sufficiently represent the composite resource entry.

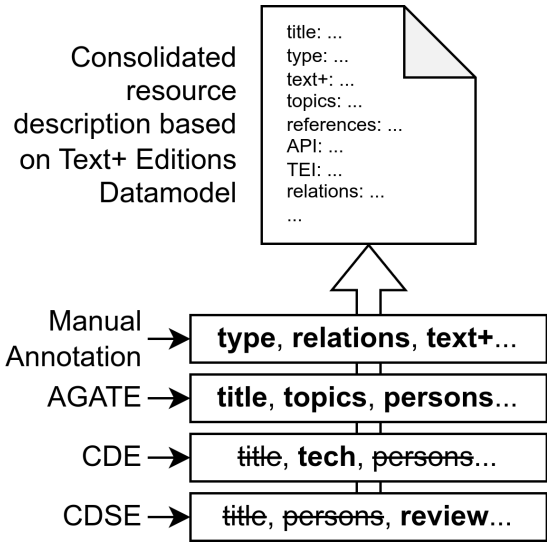

Figure 7: Exemplary description layering in editions (Gradl et al., 2024, p. 159)

4.3 Contextualisation through Cross-Referencing and Authority Data

An essential feature of the Text+ Registry is the contextualisation of individual resource descriptions through systematic cross-referencing and alignment with authoritative datasets. Registry entries are enriched not only through mutual interlinking but also by referencing external authority records. At present, the registry is connected to Wikidata and the GND for identifying and referencing entities such as persons, institutions and locations. Further examples of using controlled vocabularies and authority data includes license annotations based on the SPDX license list, as well as referencing of software utilised in editorial projects via the Research Software Directory (RSD) and the SSH Open Marketplace. Further developments will iteratively integrate additional authority data sources to overcome existing contextual limitations, such as national biases inherent in national library authority files. Planned extensions include connecting to ORCID, providing comprehensive identification of research personnel, and integrating the British National Bibliography (BNB) to expand the contextual coverage of the registry, particularly within British scholarly and cultural contexts. Overall, the integration of authoritative sources enables the creation of a network of interconnected resources and entities, significantly improving semantic coherence, discoverability, and the contextual understanding of resources.

Because both resources and their associated contexts can be searched within the registry and accessed via APIs, the registry effectively serves as a bridging component towards semantically enabled infrastructures and knowledge graph-based projects. Currently, an initial attempt at establishing these interconnections is performed automatically during external data ingestion, using matching algorithms and a comprehensive search system that integrates data from

multiple authority sources (Jegan et al., 2023). However, due to the inherent heterogeneity and variable quality of source metadata, contextualisation often remains a predominantly manual curation task, demanding expert knowledge for validation, correction, and meaningful expansion beyond automated suggestions. Future developments will aim at refining automated matching strategies and thereby reducing manual efforts without compromising data quality or accuracy. Explicit referencing of authority data and standardised identifiers facilitate alignment and interoperability with external graphs, thereby enhancing data discoverability, consistency, and integration potential. Ultimately, we anticipate that these linking practices will transform the registry into an interconnected hub, laying a robust foundation for advanced queries, reasoning capabilities, and comprehensive knowledge exploration.

4.4 Combining Data Sources for Enhanced Resource Descriptions

Data ingestion from external catalogues and manual curation efforts form the central pillars for expanding and enriching the resource landscape of the Text+ Registry. While incorporating additional external sources primarily expands the breadth of available resources, combining ingested descriptions fundamentally enhances metadata quality. Due to varying perspectives of individual sources, partial descriptions frequently complement each other, resulting in richer, composite resource metadata, as exemplified by the fields of the entry illustrated in figure 8.

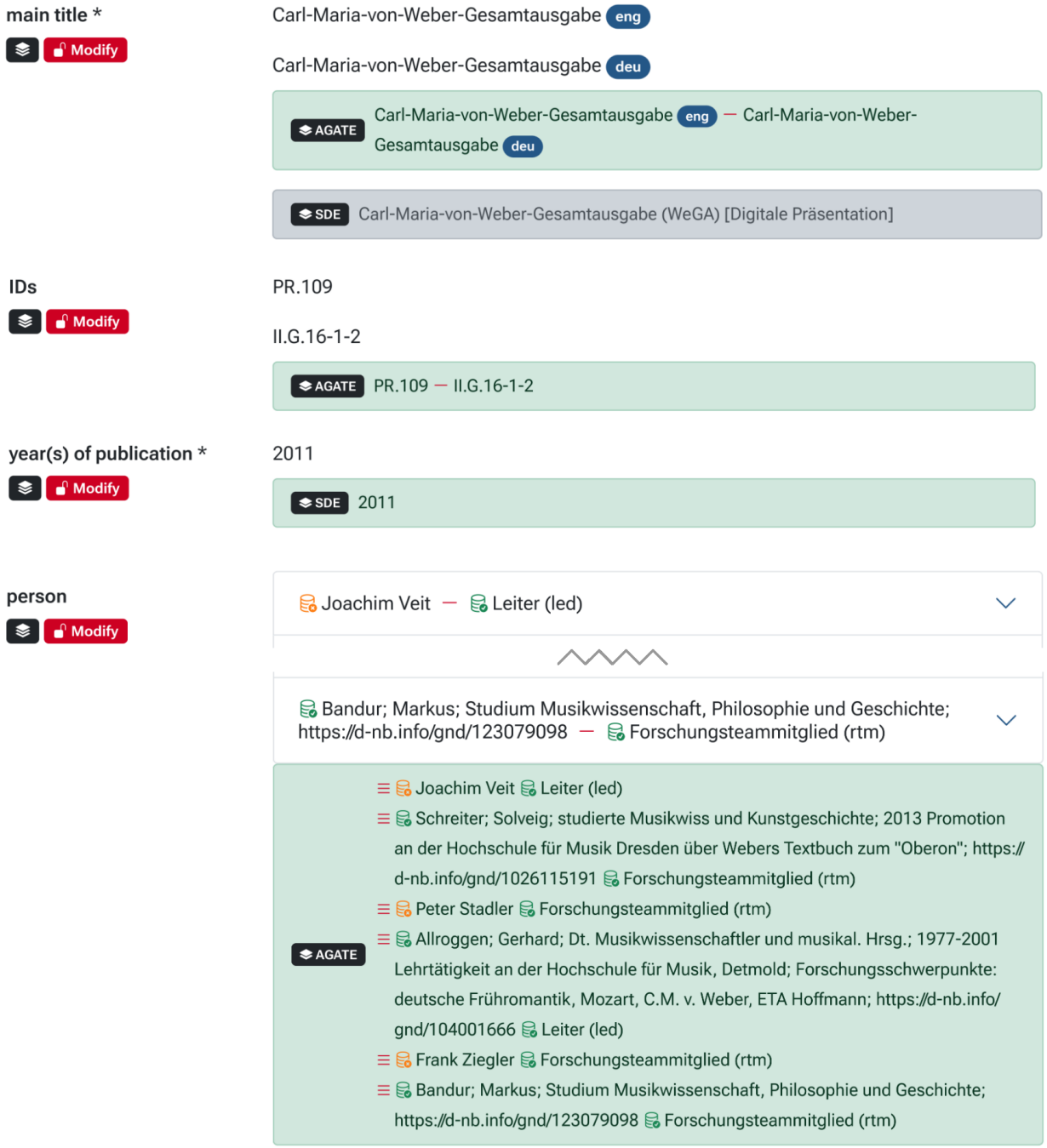

Figure 8: Fields in the Text+ Registry entry combined from multiple sources

For instance, regarding the main title field, metadata from AGATE is prioritised over data from the CDSE due to AGATE's multilingual coverage (English and German) and its broader scope, while the CDSE entry is limited to the digital presentation aspect. For the presented example, fields such as identifiers, publication year(s), and associated persons are filled from a single authoritative source. Default rules for combining data from multiple sources apply to fields such as title, description, and citation, which are populated from the highest-priority source available.

In contrast, identifiers, publication year(s), persons, and keywords aggregate data from all relevant sources, removing duplicates to present a coherent composite view.

This approach effectively balances completeness and clarity, ensuring comprehensive and authoritative resource descriptions, while still providing flexibility through manual curation to adjust priorities or overwrite values as required by expert curators.

5. Accessing the Registry with a focus on Search

The Text+ Registry offers a range of access mechanisms tailored to different user needs and technical contexts. These include both human-facing interfaces and machine-actionable endpoints for integrating registry data into external systems or workflows. While the registry supports multiple forms of interaction—including data export, harvesting protocols, and metadata transformation services—this section focuses specifically on its search capabilities, which are central to resource discovery. By examining both the web-based user interface and the structured API, we highlight how the registry enables flexible and powerful exploration of its contents.

5.1 Exploring the Registry via its Search Capabilities

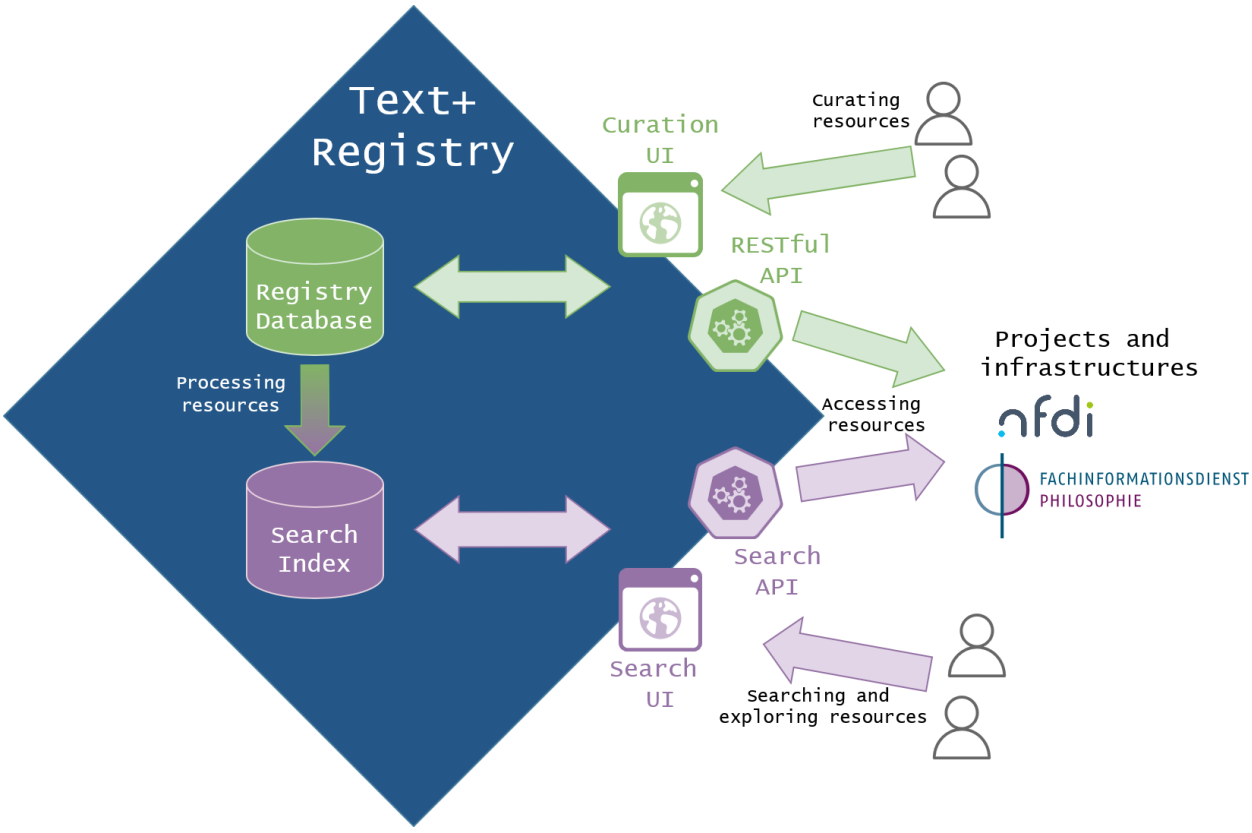

Figure 9: The endpoints of the Text+ Registry

To enhance the findability and facilitate the discoverability, the Text+ Registry offers a dedicated Search API and a corresponding search user interface – both currently under active development. Figure 9 illustrates the endpoints available with the Text+ Registry. The search endpoints retrieve the resource metadata, which are first processed and provided for search focused access. The API provides streamlined access to this diverse metadata of text- and language-based resources and offers a wide range of configurable search functionalities. These can be tailored to meet different use cases, promoting reuse across multiple application contexts, from research tools to external infrastructures. Simultaneously, the web application is designed not only as a tool for retrieving specific resources from a centralised index, but also as an interface that empowers users to navigate and explore the interconnected landscape of registry metadata, enabling both targeted queries and serendipitous discovery.

To create a unified and searchable index to support these capabilities, the imported and curated resources are processed into a central index optimised for search and query functionalities. The individual resource categories – collections, editions, and lexical resources – adhere to different data models. The import process preserves the heterogeneity of the metadata, ensuring that domain-specific information is retained in its full-depth. This enables refined, domain-specific faceted search capabilities, allowing users to filter by attributes such as provenance, disciplinary setting, or language. To enable broader interoperability and support cross-domain queries – including comprehensive filtering options – all records are additionally transformed into the DataCite metadata schema. Initially merely considered as one of the primary formats for ensuring interoperability, DataCite has proven suitable for the implementation of comprehensive search capabilities over the heterogeneous data set of the registry. The DataCite format is stored alongside the original domain-specific metadata, ensuring each resource can be queried both from a detailed, domain-focused and broad and cross-domain perspective. Provenance-aware metadata layers are consolidated according to their predefined or curated prioritisation order into a flattened representation. This reduces complexity in the search database while preserving curated edits and ensuring that enrichment, corrections, and modifications are fully reflected in the integrated, high-quality resource descriptions.
The resulting data is further refined to support **advanced search and discovery functionalities**, including but not limited to:

- Autocomplete capabilities enable immediate lookup of resources;
- Dynamic query term suggestions to help users refine their searches;
- Multilingual support, ensuring that search results remain relevant across languages. For example, the system decomposes long German compound words - breaking “Naturwissenschaft” into “Natur” and “Wissenschaft” – thus enabling retrieval when users search for either the individual components or the full term. Currently, the languages English, German, Spanish and French are considered
- Additionally, the interconnected nature of the indexed data is leveraged to generate interactive graph visualisations, revealing relationships among related resources and entities. This graphical representation not only highlights intriguing connections but also encourages users to explore connected entities inside the Text+ Registry.

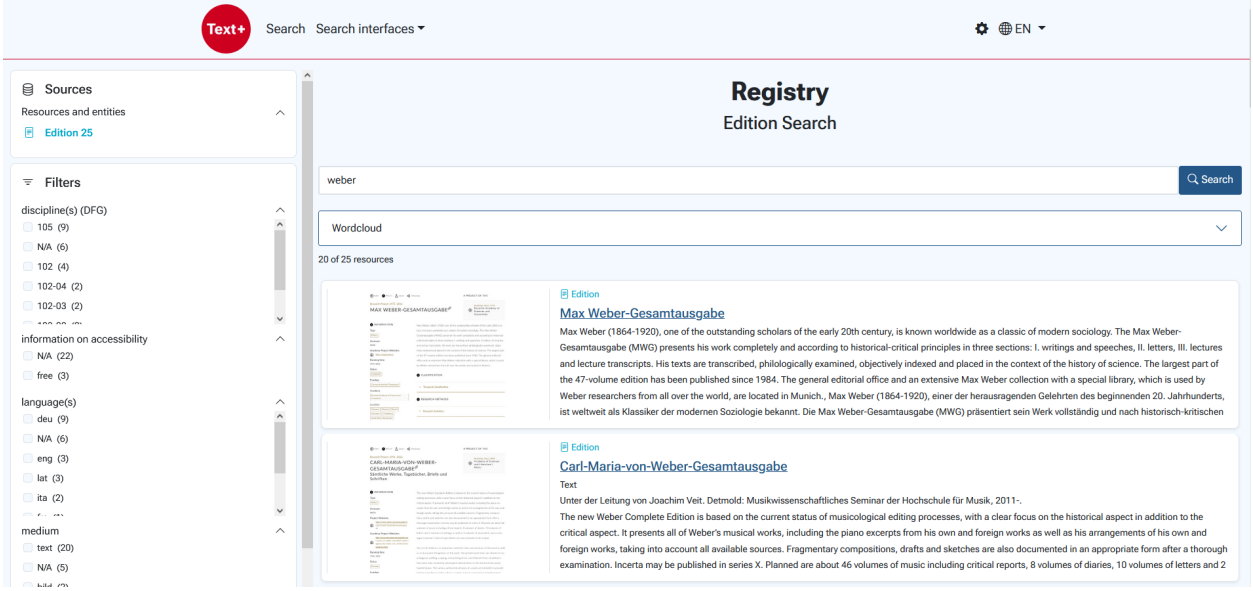

Figure 10: The Search user interface for editions

The Search API features multiple **configurable search settings** that fully exploit the unique data structures of each resource type, enabling in-depth, faceted search capabilities. This approach allows for both granular, domain-focused retrieval and broad, cross-domain queries – functionality, which is also accessible in terms of an intuitive user interface (see figure 10) that serves as a central entry point to the navigation and search capabilities of the registry. Furthermore, the system is designed for dynamic adaptability, making it straightforward to incorporate additional data formats and accommodate future projects in diverse contexts. This flexibility not only enriches the search web application, but also paves the way for the API's application in other research environments. An example is outlined in the following chapter.

5.2 Reusing the Search API

As part of the ongoing cooperation between SISs[10] and Text+, both parties not only exchange insights into each other's work in regular meetings and prepare joint papers and workshops, but also take concrete measures to interconnect their services. The SIS Philosophy integrates search results from the Text+ Registry in its service PhilFinder. In this use case, the SIS uses the registry's search API to find editions leveraging authority file identifiers.

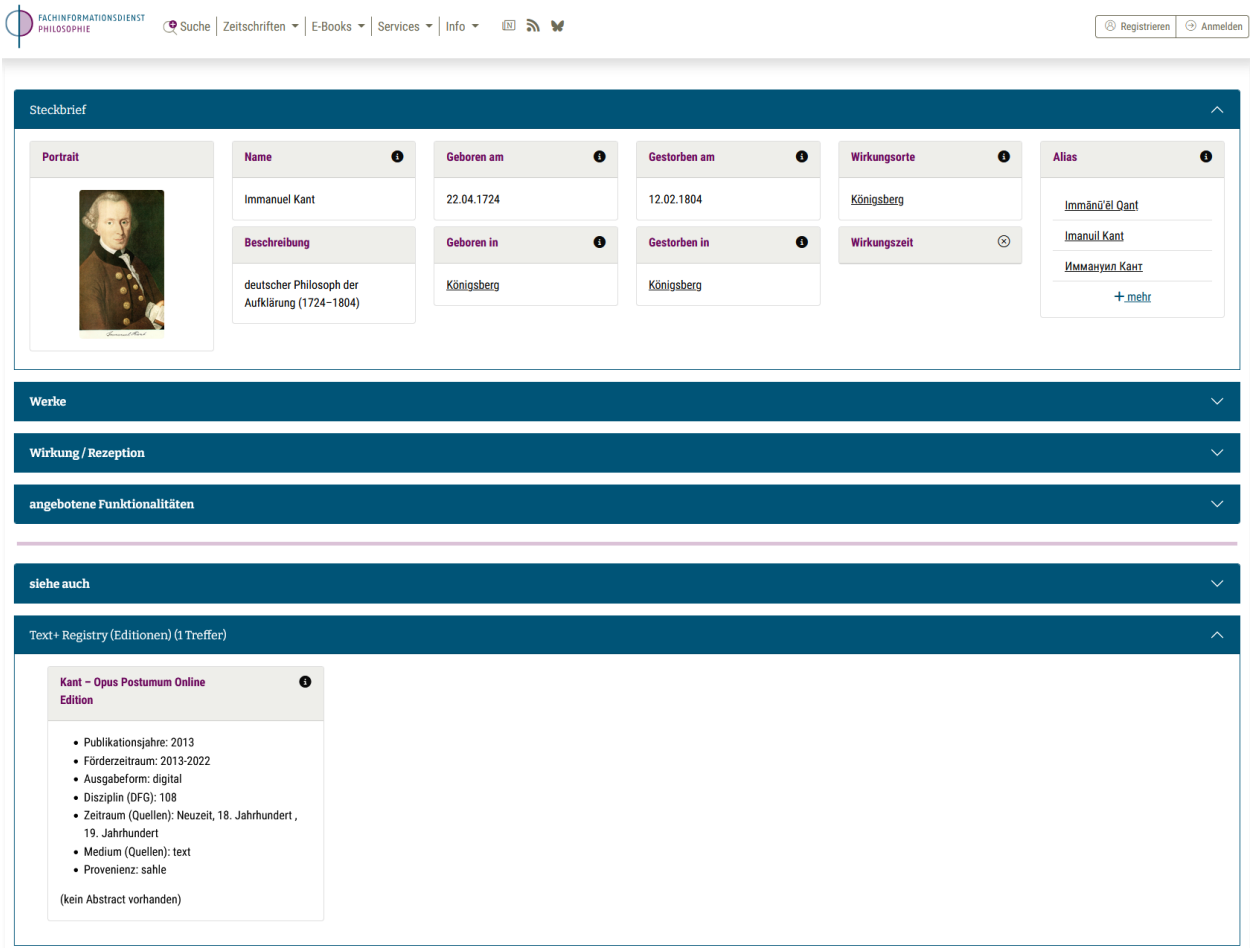

Figure 11: Immanuel Kant as shown in the PhilFinder

The PhilFinder is meant to be an entry point for research in philosophy by giving researchers shortcuts to various resources. It is under ongoing development, currently in the SIS's second funding phase and part of the application for the third, and is meant to add more and more resources over time and to become a hub for the SIS's service portfolio. It heavily relies on the LOD and FAIR principles to integrate these resources - like general openness, accessibility and shared identifiers. One of these resources is the Text+ API, which is used to deliver editions associated with a person (as an author, e.g. Immanuel Kant) to the users. The PhilFinder uses Wikidata as a LOD hub, leveraging its identifier properties to link to (or request data from) the same entities in other data and knowledge bases (cf. Zbw.eu, 2017). In this way, the PhilFinder receives the GND identifier for e.g., Immanuel Kant (GND-ID: 118559796).

Therefore, the SIS Philosophy took advantage of the possibility to enrich the data by a vertical layer and add the aforementioned identifiers to editions belonging to the domain of philosophy (via filtering for the DFG discipline). This was done manually by identifying philosophical editions via the filtering first. In the next step, either the edition's title contained the sources' author or the project website had to be searched. The approach was feasible for most of the around 80

editions present so far. The PhilFinder is now able to find editions via the registry's search API and links back to the corresponding entries. For the future there should be ways to add such identifiers automatically for example through reconciliation via OpenRefine or similar tools, or implementing union catalogues as resources for the Text+ Registry as well. The SIS also prepares a prompt to its user base for philosophical editions, trying to collect such data in a crowd sourced approach.

```
POST https://registry.text-plus.org/api/search/edition
{
 "query": "https://d-nb.info/gnd/118559796"
}
```

Listing 1: Simple HTTP POST request

The query (Listing 1) is a simple HTTP POST request with a body that specifies the search string. The uniqueness of the identifier provides high precision.
Listing 2 shows how the query can be further refined by specifying the searched indices and attributes to be filtered.[11] The following request body contains an expanded example, only searching in the editions index and filtering for resources in German, that are provided by the CDSE.

```
POST https://registry.text-plus.org/api/search/edition
{
 "query": "https://d-nb.info/gnd/118559796",
 "indices": ["registry_edition"],
 "filter": {"properties.languages.@reference": ["deu"],
          "properties.provenance.@reference": ["sahle"]}
}
```

Listing 2: HTTP POST request with additional query options

```
{
 "results": {
  "totalHits": {
   "value": 1,
```

```
    "relation": "EQUAL_TO"
  },
  "hits": [
   {
    "id": "4e1722d3-d85f-43ad-9afa-6ddda34087b4",
    "index": "registry_edition",
    "entityId": "4e1722d3-d85f-43ad-9afa-6ddda34087b4",
    "published": true,
    "title": [
     {
      "@value": "Kant – Opus\t\t\tPostumum Online Edition",
      "@lang": "deu"
     }
    ]
    ...
   }
  ]
}
```


Listing 3: Single result for the GND ID associated to Immanuel Kant

The single result is the edition “Kant - Opus Postumum Online Edition”. This proves that using such authority file identifiers ensures a high precision, but the recall is rather low. Also adding these identifiers may not be possible in every case. Using the names of entities could increase the recall, but comes at the cost of a lower precision.

When querying the Search API using either the GND or Wikidata identifier for William Shakespeare, no associated records are currently returned. In use-cases that do not need to rely on precision-focused queries, a broader search request can increase the recall and yield search results. When using the comprehensive Datacite based search[12] with the query term “Shakespeare”, the registry finds the three relevant editions “The Shakespeare Quartos Archive”, “The (New) Internet Shakespeare Editions”, and “Leipziger Ausgabe der Werke von Felix Mendelssohn Bartholdy” (which among others includes the “Music for a Midsummer Night's Dream by Shakespeare”), but also returns the “The Arden Shakespeare CD-ROM”, which is not relevant for our assumed information need. Since the query is in this case not limited to the editions domain, a total of eight poem text corpora are among the results, which the University of Tübingen has collected as part of the “Interpretability in Context” project. Most of the returned corpora are titled as “Poem corpus Dickinson / Donne / Shakespeare” and the

concrete relevance for our assumed intent again needs to be verified qualitatively by the searching user.

```
POST https://registry.text-plus.org/api/search/datacite
{
 "query": "Shakespeare"
}
```

Listing 4: Full-text query with the integrated DataCite based search

However, the Search API does not only provide a full-text search that retrieves documents based on a single, otherwise unspecified query term. Queries can focus on specific fields e.g., to search for a relevant name or identifier only in originator or person fields, but not in titles or descriptions. In addition, Boolean operators and other features are in place to allow queries such as “William AND Shakespeare”, requiring both parts of the name to occur in the document. Beyond these refinements to search functionality, improvements to the Search UI will further enhance user experience and discoverability in the Text+ Registry. This includes an improved presentation of metadata and their connections between resources, using interactive graph visualisations that facilitate intuitive exploration and navigation. By strengthening both the API’s search capabilities and the user interface, the system continues to evolve in response to the requirements of the registry user communities.

6. Conclusions and Future Plans

The Text+ Registry serves as a foundational component for improving the FAIRness of resource descriptions within text- and language-based research. Its hybrid architecture enables both the reuse of metadata from authoritative external sources and the manual curation of new and ingested entries, addressing the need for both consistency and domain-specific detail. Through provenance-aware metadata layering, flexible data models, and tailored search capabilities, the registry offers a robust framework for organising, contextualising, and discovering a diverse range of scholarly resources. An example of an implementation of the Search API is presented in the use case of the SIS Philosophy’s PhilFinder, where it plays an important role in providing researchers with relevant scholarly editions.

Looking ahead, the registry will continue to expand its integration of external authority data, such as ORCID and the British National Bibliography, to enrich contextual relationships and reduce national or disciplinary silos. Further development will also focus on enhancing automation in data linking and enrichment, improving semantic interoperability–positioning the registry as a bridge to external knowledge graphs, Linked Data networks, and other research infrastructures within the NFDI and beyond. To fully realise this potential, the continued integration of additional catalogues—including bibliographic, archival, and software registries—is essential. These integrations will not only enhance the contextual depth of registry entries but

also contribute to a more connected, multilingual, and semantically navigable research landscape.

In addition, continued emphasis will be placed on usability—both through the web interfaces and the APIs—to ensure that the registry evolves to become a sustainable, extensible, and accessible infrastructure for the Digital Humanities. Besides technical considerations and enhancements, here also further community engagement plays a vital role in the establishment of the Text+ Registry as an accepted and widely used research tool and meta catalogue.

Acknowledgements

Special thanks are due to the Herzog August Bibliothek Wolfenbüttel, which provided the financial means for Daniela Schulz to attend the event and give the presentation that resulted in this paper.

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.

References

Cugliana, E., Gengnagel, T. and Hensen, K. (2022) Das Portfolio der TA Editions. Text+ Plenary 2022 (TextPlusPlenary), Zenodo. doi: <https://doi.org/10.5281/ZENODO.7244299>.

Gödel, M., Klappenbach, L., Sander, R. and Schnöpf, M. (2024) Wer sind die Herausgeber:innen Digitaler Editionen? Eine Untersuchung zur Repräsentation von Digital Humanities-Wissenschaftler:innen. DHd 2024 Quo Vadis DH (DHd2024), Zenodo. doi: <https://doi.org/10.5281/ZENODO.10706125>.

Gradl, T., Kudella, C., Lordick, H. and Schulz, D. (2024) Towards a Registry for Digital Resources – The Text+ Registry for Editions. *Datenbank-Spektrum*, 24(2), pp.151–160. doi: <https://doi.org/10.1007/s13222-024-00479-0>.

Jegan, R. *et al.* (2023) 'Integrating Access to Authority Data for Improved Interoperability of Research Data in the Digital Humanities'. doi: <https://doi.org/10.18420/BTW2023-54>.

M1 der Task Area Editions in Text+ (2024) Die Editionsregistry (ERY) und das zugrundeliegende Datenmodell – Dokumentation und Spezifikation. doi: <https://doi.org/10.5281/ZENODO.13379995>.

Measure 1 der Task Area Editions (2024) Datenmodell Editionenregistry (Text+). doi: <https://doi.org/10.5281/ZENODO.12799883>.

Sahle, P. (2013) *Das typografische Erbe. Digitale Editionsformen : zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*. Norderstedt: Book on Demand (Schriften des Instituts für Dokumentologie und Editorik, 7).

Schulz, D., Fisseni, B. and Sendler, S. (2021) EdMA – Eine Matrix zur Erfassung und Kategorisierung digitaler Editionen. doi:<https://doi.org/10.14618/IDS-PUB-10501>.

Zbw.eu. (2017) *Wikidata as authority linking hub: Connecting RePEc and GND researcher identifiers* | ZBW Labs. [online] Available at: <https://zbw.eu/labs/en/blog/wikidata-as-authority-linking-hub-connecting-repec-and-gnd-researcher-identifiers.htm>l [Accessed 2 Jun. 2025].

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.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. and Gonzalez-Beltran, A. (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, [online] 3(1). doi: <https://doi.org/10.1038/sdata.2016.18>.

Wuttke, U., Ott, C., Adrian, D. and Worthington, S. (2017) AGATE: Concept for a European Academies Internet Gateway for the Humanities and Social Sciences. doi <https://doi.org/10.5281/ZENODO.815916>.

About the Authors

Leon Fruth is a researcher in the Media Informatics Group at the University of Bamberg, Germany working on the search functionalities of the Text+ Registry.

Nils Geißler is a researcher at Cologne University, Germany. There, he holds a matrix position working both for the Text+ consortium in the data domain Editions and for the SIS Philosophy. He is mainly responsible for the development of the PhilFinder service.

Tobias Gradl is the main developer of the Text+ Registry working for the Media Informatics Group at the University of Bamberg, Germany.

Daniela Schulz is a researcher at Herzog August Library Wolfenbüttel, Germany working in the data domain editions of the Text+ consortium.

---

[1] AGATE is the research information system of the German union of Academies, an umbrella organisation of eight German academies of sciences and humanities, that act as funding bodies for often long term research projects. Cf. <https://www.akademienunion.de/en/research/agate>. (Accessed: 2 June 2025)

[2] In addition to the aforementioned existing catalogues, the preliminary work from the previous CLARIAH project, in which an edition matrix for describing digital editions was created, should be noted here in particular. Cf. Schulz, Fisseni and Sendler, 2021.

[3] The editions belonging to the portfolio were first recorded in 2022 by means of a survey; an update was carried out in 2023. An overview of the initially recorded portfolio with some analyses was presented in a poster at the 1st Text+ Plenary 2022. Cf. Cugliana et al., 2022.

[4] SISs are funded by the DFG and aim at providing scholars with appropriate specialised literature, information, and services that are relevant to their research.
[5] For the specification of the data model and further information (in German) cf. M1 der Task Area Editions in Text+, 2024.
[6] Cf. Gödel et al., 2024.
[7] One problem with the DFG subject classification system is the lack of representation of the so-called 'small subjects'. Furthermore, the system is subject to regular changes, so that its use as a taxonomy is highly problematic. For the current version see <https://www.dfg.de/resource/blob/331950/85717c3edb9ea8bd453d5110849865d3/fachsystematik-2024-2028-en-data.pdf> (Accessed: 2 June 2025). For the BK, among others the SISs and Text+ currently cooperate in its translation to English. Cf. <https://wiki.k10plus.de/pages/viewpage.action?pageId=740393030>. (Accessed: 2 June 2025)
[8] <https://dig-ed-cat.acdh.oeaw.ac.at/documentation.html>. (Accessed: 2 June 2025)
[9] <https://www.digitale-edition.de/exist/apps/editions-browser/about.html>. (Accessed: 2 June 2025)
[10] The SISs are funded by a special funding line of the DFG under the LIS (Literary Information Systems) line. There are over 60 SISs that are clustered mostly by discipline and that are organised in an overarching working group (AG FID). There are four possible funding phases before entering the so-called FIDplus programme, which is currently developed by the DFG.
[11] More examples on how to use the API: <https://registry.text-plus.org/docs/docs/category/api>. (Accessed: 2 June 2025)
[12] For reproduction in the user interface: <https://registry.text-plus.org/default/search?query=Shakespeare>. (Accessed: 2 June 2025)