Text+: A National Hub including Legacy Language Data

Florian Barth[*], **Christoph Draxler**[†], **Jennifer Ecker**[**], **Stefan Fischer**[+]
Philippe Genêt[‡], **Alina Hemmer**[§], **Timm Lehmberg**[¶], **Thorsten Trippel**[‖] [**]
Andreas Witt[**], **Arden Zimmermann**[‡], **and Claus Zinn**[‖]

[*]University of Göttingen
Papendiek 14, D-37073 Göttingen
florian.barth@uni-goettingen.de

[†]Bavarian Speech Archive (BAS)
Schellingstr. 3 / II, D-80799 München
draxler@phonetik.uni-muenchen.de

[+]Saarland University
Campus A2 2, D-66123 Saarbrücken
stefan.fischer@uni-saarland.de

[‡]German National Library
Adickesallee 1, D-60322 Frankfurt am Main
{P.Genet, Ar.Zimmermann}@dnb.de

[§]University of Hamburg
Max-Brauer-Allee 60, D-22765 Hamburg
alina.hemmer@uni-hamburg.de

[¶]Academy of Sciences and Humanities in Hamburg
Edmund-Siemers-Allee 1, D-20146 Hamburg
timm.lehmberg@awhamburg.de

[‖]University of Tübingen
Keplerstraße 2, D-72074 Tübingen
{claus.zinn, thorsten.trippel}@uni-tuebingen.de

[**]Leibniz Institute of the German Language
R 5, 6-13, D-68161 Mannheim
{ecker, trippel, witt}@ids-mannheim.de

Abstract

Text+ is the German distributed research data infrastructure for literary studies, linguistics, and spoken and written language. Its resources consist of contemporary and historical literary and media texts, deeply annotated material, transcripts of spoken and sign language, and original recordings. Text+ provides access to its resources according to the FAIR guidelines: Findable due to standard-conformant metadata, Accessible with single sign-on authentication, Interoperable via open data formats, and Reproducible through web services and extensive documentation. The 30+ partners of Text+ are archives, libraries, universities, and other research institutions. The partners are autonomous, and they differ in the amount of data and processing capabilities they provide. In this paper, we describe the hub architecture of Text+, which gives users a central and FAIR point of access to research data that continues to be distributed across the Text+ partner institutions. The architecture serves as a blueprint to evolving research infrastructures that aim at maintaining (and empowering) their research data contributors.

Keywords: national infrastructure for research data management, federated data, workflows and policies

1. Problem Statement

The Text+ consortium (Hinrichs and Trippel, 2024) was created to support academic research that relies on language and text as primary data. Such data is available in huge quantities, in written and spoken modalities, including sign language corpora, in the German language or in German dialects. There are many corpora of newspapers, audio and video archives, in part transcribed, and

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8264–8275
11–16 May 2026.

data that is annotated with different linguistic layers, ranging from morphology to argument structures. Such data comes with yet other qualities which encompass issues such as data protection and licencing. Making all such data available at researchers' fingertips, given that it is distributed over 30+ different locations is no easy matter, especially as the institutions are of various types, including universities, academies, and libraries. How to provide researchers with a central point of access to find and use research data when all 30+ institutions are (and continue to be) autonomous entities? In this paper, we – from the perspective of one of Text+' data domains – describe the hub architecture of the Text+ consortium, which builds on the strength of its members. Our approach relies on the harmonization of metadata, and the clever use of interfaces to provide easy access to data and services distributed over a federation of data centres.

The phrase "The whole is greater than the sum of the parts" is attributed to Aristotles. Here, Text+ as an umbrella organization (the whole) aims at providing added value to those services provided in a distributed manner (the parts). The added value of Text+ can be described in terms of FAIR data management, because a central access point to research data increases the use of research data across all four FAIR dimensions.

2. Background

The Text+ consortium is one of the 27 consortia that form the national research data infrastructure in Germany.[1] As one of six consortia situated in the Humanities and Social Sciences, Text+ aims at providing a central access point to all researchers who use or produce language-related data for their studies.[2] Initial funding started in October 2021, with now 30+ institutions participating and contributing their research data and their expertise for managing it. The Text+ consortium, along with most of the other NFDI consortia, takes a federated approach. Rather than building a nation-wide archive to hold all language data in a single location (so that the 30+ institutions can be dissolved), it targets to be an umbrella organization, a national association that brings together the smaller organizations under a broader structure, bundling their shared interests to make research data FAIR (Wilkinson et al., 2016).

Text+ follows the tradition of previous infrastructures, such as the German CLARIN-D infrastructure, which in turn, took part in the European CLARIN project.[3] Text+ adopts the hub approach from CLARIN-D, and also builds on developments within DARIAH-DE, which is the national partner of another European project, DARIAH[4]. The hub design, however, needs to scale-up to a larger number of partner institutions, and an even more diverse set of research data. Also, the need to address the enlarged community with a unified voice becomes more ambitious.

While large-scale research infrastructures usually follow a federated, hub-based approach, there are also purely centralized bodies that offer research data management services to larger communities. The Linguistic Data Consortium (LDC) has a long tradition in servicing the linguistics community, making available a wide variety of language-related resources.[5] LDC has a centralized steering organization that has full control over the entire data lifecycle, from data generation to data provision. It manages its own data repositories and takes care of licensing, quality control and user access in a central manner.

The ELRA Language Resources Association is a central agency with the goal of making available language resources to the scientific community.[6] According to their website, its services include the identification of language resources, the promotion of the production of language resources, the production, validation and distribution of language resources, as well as the evaluation of systems, products, and tools related to language resources.

The Language Data Commons of Australia (LDaCA)[7] is a national initiative focused on the preservation, accessibility, and reuse of language data, particularly in the context of indigenous and multilingual research. It operates as part of the Australian Research Data Commons (ARDC)[8], with the mission of accelerating Australian research and innovation through the creation, analysis, and retention of high-quality data assets across research domains. LDaCA aims to create a federated archive to secure language data, with a multilingual focus spanning from Aboriginal languages, Pacific regional languages, and Australian English. LDaCA's integration within ARDC illustrates a model of national coordination that supports both disciplinary specificity and cross-sector collaboration.

In contrast to centralized infrastructures, hub-based approaches must cope with and make centrally available diverse sets of research data and services that were created outside of the hub's control, in research organizations that all have their own institutional research foci, disciplinary practices, and also legal constraints.

[1] https://www.ndfi.de
[2] https://www.text-plus.org
[3] https://www.clarin.eu
[4] https://www.dariah.eu
[5] https://www.ldc.upenn.edu
[6] https://www.elra.info
[7] https://www.ldaca.edu.au
[8] https://ardc.edu.au

3. Exemplary Research Data in Text+

The 30+ partner organizations of Text+ represent a wide set of research areas, which we compartmentalized into three distinct data domains within dedicated Task Areas (TA): Editions, Lexical Resources, and Collections.[9] The Editions TA focuses on scholarly editions, particularly historical documents and publications contextualized within their respective philological traditions. Lexical Resources encompass a wide range of lexical data, from digitized print dictionaries to complex wordnets and other structured lexical databases. The Collections TA includes diverse types of corpora and datasets generated in linguistic and literary research, such as spoken and written language corpora, questionnaires, and empirical studies in language science. This paper is situated within the Collections TA. To illustrate the wide scope of this TA, we describe eight Text+ partner institutions and their datasets to highlight the challenges Text+ needs to address.

The German National Library (DNB) is one of the largest partners in the Text+ consortium. The DNB operates under a legal mandate to collect, document, and preserve all publications produced in Germany or in the German language, as specified in national law, particularly in the *Gesetz über die Deutsche Nationalbibliothek* (DNBG). Publishers are required to deposit their works, but this obligation does not automatically include permission for redistribution or reuse within research infrastructures. Access to deposited content is therefore often constrained by copyright and licensing restrictions, and materials are received in heterogeneous formats such as EPUB and PDF, sometimes including digital rights management mechanisms.

Within Text+, the DNB contributes bibliographic and entity reference data. The DNB maintains Germany's national bibliographic metadata and the authority file *Gemeinsame Normdatei* (GND), a system for standardized entity reference data covering persons, organizations, subjects, and places. GND data are provided in MARC 21 and RDA formats, and are also available as Linked Open Data through RDF serializations and a public SPARQL endpoint.

All DNB bibliographic and entity reference metadata are published under CC0[10], allowing reuse in other catalogues, repositories, and research infrastructures. This open and standards-based approach facilitates the alignment of Text+ resources – such as corpora, lexicographic data, or digital collections – with persistent identifiers and GND entities, enabling interoperability without imposing a centralized repository model.

The Leibniz Institute for the German Language (IDS) curates the German Reference Corpus (DeReKo), the largest linguistically motivated collection of German texts. These resources are processed using TEI-based formats, with workflows that have been refined since the 1960s. The institute employs XML-based structures such as the TEI P5 (TEI Consortium, 2025) subset known as I5 (Lüngen and Sperberg-McQueen, 2012), which allows for syntactic validation against XML schemas.

With its expertise in corpus linguistics and language resource management, IDS is also a partner in the CLARIN-D infrastructure and contributes to the European network of national language institutions (EFNIL)[11]. The repository system used at the IDS is based on Invenio (Saleh et al., 2025) and supports metadata standards such as CMDI (ISO 24622-1; ISO 24622-2) and DataCite[12], and persistent identifiers following ISO 24619 (ISO 24619). The use of these standards ensures the interoperability and long-term accessibility of the resources.

The IDS repository is also open to external data contributions, particularly those created within Germany or involving the German language. Resources are accepted if they conform to the IDS' archival formats such as TEI P5 based I5 or ISO 24624 (ISO 24624) for spoken language data.

The IDS provides access to its curated corpora, including DeReKo, under a rights and access management framework designed to balance openness with legal and ethical constraints. While some of the raw data cannot be made publicly available due to licensing restrictions, most resources are accessible free of charge to the academic research community after authentication via login[13]. Most corpora are available under a Query Analysis Only, non commercial (QAO-NC) licence. This controlled access model ensures compliance with third-party rights while enabling broad scholarly use.

IDS recognizes that access must extend beyond academia. As part of its commitment to FAIR principles and inclusive research, IDS implements mechanisms to support educational and non-commercial use, including by teachers and learners outside of traditional academic research settings.

The University of Tübingen (EKUT) hosts the CLARIN-D Repository, a CoreTrustSeal-certified trustworthy data repository. The repository hosts a

[9] The three data domains are complemented with an overarching TA Infrastructure/Operations. Note that a partner institution is part of at least one TA.

[10] https://www.dnb.de/EN/Professionell/Metadatendienste/metadatendienste_node.html

[11] https://efnil.org

[12] https://datacite.org

[13] https://www.ids-mannheim.de/en/digspra/pb-s1/projects/corpus-development/verfuegbarkeit

large variety of research data, most of which stem from two Collaborative Research Centres (CRC) in linguistics, and which need to be archived for a minimum duration of ten years as means of good scientific practice. All data are described with both CMDI-based metadata and DataCite. Each dataset is addressed with a DOI-based persistent identifier. All metadata is made available via an OAI-PMH endpoint and is indexed, for instance, by the CLARIN Virtual Language Observatory (Van Uytvanck et al., 2010), and the Text+ registry, see below. Among the noteworthy datasets is the Tübingen Treebank of Written German (TüBa-D/Z), a collection of articles from the daily newspaper, "die tageszeitung", which have been automatically annotated with clause structure, topological fields, and chunks, in addition to more low level annotation including parts of speech and morphological ambiguity classes; TüBa-D/DP, a machine-annotated dependency treebank of German, offering high-quality syntactic annotations for a huge amount of contemporary German text; and TüBa-D/W, a large treebank of modern written German, which is based on Wikipedia text consisting of 36.1 million sentences (615 million tokens) in CoNLL-X format.

EKUT also hosts GermaNet, the largest lexical-semantic net for German; it relates German nouns, verbs, and adjectives semantically by grouping lexical units that express the same concept into synsets and by defining semantic relations between these synsets (Hamp and Feldweg, 1997). Moreover, EKUT gives access to all UD treebanks[14] via TüNDRA (Martens, 2013), a web-based tool for treebank research.

The EKUT research data comes with a diverse set of legal requirements. GermaNet, for instance, requires a usage licence but is free for academic use; data based on newspaper corpora is protected and can only be distributed as a derived format, see below. The UD treebanks are all public, but there are also significant CRC data, which is not public or currently under an embargo period.

Göttingen State and University Library (SUB), one of the largest academic libraries in Germany, offers comprehensive services for the archiving and provision of research data in the humanities. In Text+, SUB functions as a data centre, particularly for text-based research data. SUB operates both the DARIAH-DE and the TextGrid Repository, both of which obtained the CoreTrustSeal in 2024.

The *DARIAH-DE Repository* is a digital long-term archive for research data in the humanities, operated by the SUB since 2017. Each archived object is assigned a DataCite DOI, which ensures that it remains permanently referenceable, citable, and accessible. The DARIAH-DE Repository is ideally suited for generic research data and collections – there are no restrictions on the types of data accepted. The *TextGrid Repository* is a long-term archive for research data in the humanities and specialized for XML/TEI-encoded texts. It provides an extensive, searchable, and reusable collection of texts and images. For researchers, it offers a sustainable, permanent, and secure platform for the citable publication of digital texts connected to several options, including transformation, description, searches, and connection to other resources and tools.

New features like customisable project pages and project-specific facets have been added during the course of Text+. Also, the recently introduced *Fluffy Publication Workflow* further streamlines the publication process in the TextGrid Repository (Veentjer et al., 2025) by integrating tools such as TEI, XPath, Git, and Jupyter Notebooks. Users can enrich and validate metadata without altering the original data files, thus improving the overall data quality and compliance with the FAIR principles. The workflow generates the necessary metadata files for the publication and supports seamless upload to TextGrid via a Jupyter-based interface. This approach lowers technical barriers for researchers while ensuring sustainable, high-quality, and citable publication of textual research data.

The Bavarian Archive for Speech Signals (BAS) is hosted by the Institute of Phonetics and Speech Processing at LMU Munich. BAS maintains a CoreTrustSeal certified repository of spoken language corpora and provides a series of spoken language processing web services. The repository contains 50+ corpora, mainly of spoken German, targeted at speech technology development and evaluation, and fundamental research[15]. The repository is based on CMDI-compliant metadata, is connected to the Text+ registry, and regularly harvested by other academic and research search engines.

The BAS web services[16] (Kisler et al., 2017)) provide an easy-to-use graphical interface to complex speech processing tasks to academic users worldwide. Time-consuming or heavy processing tasks such as automatic speech recognition or word- and segment-level labelling and segmentation in more than 50 languages can be accessed securely from any client computer, or via API calls. This allows transcription editors such as ELAN (Sloetjes et al., 2007), EXMARALDA (Schmidt, 2012) or Octra (Draxler and Pömp, 2022), or custom scripts, to execute web service calls in the background to

[14] https://universaldependencies.org

[15] https://clarin.phonetik.uni-muenchen.de/BASRepository

[16] https://clarin.phonetik.uni-muenchen.de/BASWebServices

enhance user experience. In the last three years, the web services have processed approximately 1 million media files per year, with access from more than 70 countries.

BAS has also developed speech processing software, which have become de facto standards: SpeechRecorder (Draxler and Jänsch, 2004) and WikiSpeech (Draxler and Jänsch, 2008) for scripted audio recordings, WebMAUS (Kisler et al., 2012) for automatic segmentation, EMU/SDMS (Winkelmann et al., 2017; Jochim, 2017) for corpus management and statistical analysis of spoken language data, and the transcription management system Octra Backend (Draxler and Pömp, 2024).

University of Hamburg (UHH), Hamburg Centre for Speech Corpora (HZSK) provides access to a wide range of language corpora via the Research Data Repository of the University of Hamburg. While a particular focus is placed on spoken language data and linguistic fieldwork, the HZSK collections also holds a broad spectrum of linguistic resources, covering research areas such as language variation, language acquisition, multilingualism, as well as institutional and medical communication. In addition to its role as a Text+ data centre, the HZSK is accredited as a CLARIN C-Centre and is a member of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation.

A specific emphasis lies on data from the domain of medical communication. The DiK corpus (Bührig and Meyer, 2009), for instance, consists of audio recordings, complemented by linguistic transcriptions and annotations for various aspects of doctor-patient communication in hospitals. Included are both monolingual conversations in German, Portuguese, and Turkish as well as bilingual interactions where interpretations between some of these languages take place.

Given the highly sensitive nature of this data – particularly due to the presence of personal health information and the vulnerability of the persons involved – access to the DiK corpus is subject to strict rights management. Access is granted exclusively to eligible researchers upon individual application.

HZSK also provides numerous corpora that are made available under open licences. HZSK-Community corpora are prepared using established standards for the transcription and annotation of spoken language, most prominently the EXMARaLDA XML-based TEI format.

HZSK corpora are curated and published within a dedicated community in the University of Hamburg's FDR. The FDR is a permanent, institutionally supported service based on Invenio, supporting persistent identifiers, CMDI and DataCite, and an OAI-PMH endpoint for metadata harvesting. – External data contributions can also be curated and published through the HZSK community, provided they meet requirements regarding quality, documentation, ethical clearance, and relevance to the centre's thematic scope (Bührig, 2025).

CLARIND-UdS is the designated repository for language resources at Saarland University. As implied by its name, the data centre was already part of the German CLARIN-D initiative. The centre is operated by the Department of Language Science and Technology, which conducts research in computational linguistics, corpus linguistics, phonetics, psycholinguistics, and translation. Due to its research profile, the centre's archival policy prioritizes the preservation of non-German and multilingual text corpora. The repository prefers resources in verticalized text format (VRT), which are compatible with the Corpus Workbench (CWB) and CQPweb tools used by many corpus linguists. Consequently, the centre's backend for the Text+ federated content search (FCS) is also based on the CWB platform. The repository runs on Fedora Commons and provides metadata in two standardized formats: Dublin Core and CMDI. Archived resources are assigned persistent identifiers via the Handle System[17]. The repository was certified by CoreTrustSeal in 2024.

The Academy of Sciences and Humanities in Hamburg (AdWHH) contributes multimodal and multilingual language resources to the Text+ consortium, with particular expertise in sign language data and endangered language documentation. Using its own discipline-specific interface along with the Text+ specific interfaces, the AdWHH provides access to unique corpus collections, focusing on resource types that require specialized multi-layered annotation and preservation approaches. A flagship resource is the *German Sign Language Corpus (DGS-Korpus)*[18], which represents one of the largest systematically collected and annotated sign language corpora worldwide. This corpus contains video recordings from over 330 deaf signers across Germany, capturing regional variations and generational differences in German Sign Language. The data includes detailed multi-tier annotations covering manual signs, non-manual components, and German translations, making it an invaluable resource for sign language linguistics, lexicography, and computational modeling. Beyond this, AdWHH also provides unique language documentation, i.e., deeply annotated spoken language data from endangered languages of Northern Eurasia.

Summary The eight partners of the task area Collection contribute a rich and diverse set of data

[17] https://hdl.handle.net

[18] https://www.sign-lang.uni-hamburg.de/dgs-korpus

Institution	Type of data	type of metadata	Access & Rights
DNB	All publications in Germany, or in German language, GND data	MARC-21, RDA, RDF	Restricted, bibliographic data and GND data: CC-0
IDS	Collections of German Text, DeReKo	CMDI, DataCite	Restricted, Academic Licence, QAO-NC
EKUT	CRC data, GermaNet, Tüba/DZ, Tüba/DP, UD treebanks	CMDI, Dublin Core	Restricted, Academic Licence, CC-BY
SUB	Any media (text, image, audio), XML/TEI-encoded texts	TextGrid metadata schema, Dublin Core	Open, user defined license (recommended: Academic License, CC BY-NC-SA 4.0)
BAS	Spoken language data	CMDI	Restricted
UHH	Spoken language data, fieldwork data, medical texts	CMDI, DataCite	From open to highly restricted
UdS	Written corpora (non-German and multilingual)	CMDI, Dublin Core	Open, restricted
AdWHH	Spoken language data, fieldwork data, sign language data	CMDI, DataCite	Open

Figure 1: Research data, Metadata, and Access/Rights management.

and metadata, infrastructure, and expertise to the Text+ consortium. All partners maintain their own repository services, and most partners have had API endpoints in place prior to their Text+ era. Fig. 1 summarises the Text+ partners' resources.[19]

4. Text+ Approach

The Text+ umbrella organization around the 30+ partners is built on a number of pillars of which metadata harmonization is a central one. A common understanding of the metadata used, and hence, of the multitudes of existing research data, makes it possible to build central services to make all data FAIR-ly available to the Text+ community.

4.1. Harmonization of Metadata

Each Text+ partner uses their own metadata formats, often tailored to their local specializations and requirements, to describe their research data. As these are created in established workflows, it was clear from the outset, that partner organizations are going to uphold their metadata practices so that no new metadata standards could be imposed from the umbrella organization to its centre outposts. A centralised and effective search across distributed datasets, however, requires a common understanding of all metadata schemas in use.

The Text+ consortium invested considerable time to understand the metadata schemas in use and to extract their common core. Each of the three Text+ TAs has now defined a core metadata set, which is currently expressed in three Excel sheets. TA Lexical Resources has now formalized this tabular form using CMDI. This schema consists of seven components, each of which groups together one aspect of metadata such as administrative, bibliographic, technical, spatial, and legal metadata fields as well as information on the data' lifecycle. A seventh component is specific to the task area and represents, for instance, type, modality, and entry type of a lexical resource. It is planned that the six resource-independent components are also used by Collections and Editions. For Collections, the seventh component represents information about a collection's genre, collection type, disciplinary classification, and subject headings, among others. The three schemas inform the Text+ Registry, the central access point for metadata-based search.

4.2. Centralized Metadata Search

The Text+ Registry Gradl et al. (2024)) aims at making all Text+ research data *Findable* via metadata-based search.[20] The catalogue makes use of the metadata harmonization effort described earlier. Using the registry's domain modelling environment (DME), data harvested from multiple parties, using different metadata formats, are mapped onto one of the three TA-specific common data models.

For each task area, a tailored graphical user interfaces is bootstrapped with the help of the respective data model. While each interface provides users with access to the metadata of its task area via a combination of faceted and full-text search, the number and the nature of the facets change depending on the model. In the Collections Registry, *e.g.*, a facet MODALITY with values WRITTEN, SPOKEN, SIGNED, MULTIMODAL and OTHER will be shown.

In addition to searching within each individual

[19] Note that a Text+ partner must take part in at least one task area. There are partners, however, which contribute to two task areas given that the data and services they host are associated with multiple TAs.

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

area, a unified interface is available that enables cross-domain metadata exploration, allowing users to search across all three areas simultaneously.

4.3. Federated Content Search

Content-based access to research data is essential for many use cases. Unlike metadata, which is typically standardized and curated, content data is heterogeneous in terms of type, format, and quality. These differences often reflect the workflows and practices of the originating institutions.

Many Text+ partner institutions operate mature local infrastructures and APIs that enable content search within their holdings. For example, the IDS provides access to DeReKo via the powerful search tool KorAP (Diewald et al., 2016; Diewald and Margaretha, 2016). Also, the University of Tübingen offers TüNDRA, a tool for querying syntactically annotated treebanks. These systems are tailored to specific resource types and offer rich query languages, but they are not easily transferable across different data classes.

To enable cross-institutional and cross-resource content search, the Federated Content Search (FCS) protocol was developed within the European CLARIN infrastructure (Schonefeld et al., 2014). FCS is based on the SRU/CQL standard and allows querying distributed resources through a unified interface. Within Text+, the FCS specification has been extended to support additional resource types, such as lexical databases (*e.g.*, GermaNet) and syntactically annotated corpora (*e.g.*, Universal Dependencies treebanks). These extensions make richer linguistic data accessible and searchable in a federated manner (Körner et al., 2025).

In practice, a central aggregator issues a query in the FCS query language, which is forwarded to all connected endpoints. Each endpoint translates the incoming FCS query into the native query language of its local system, executes the query, and returns the results in a standardized format. While the FCS query language represents only a subset of the expressive power of the local systems, the returned results typically include links to the original systems, allowing users to explore the full capabilities and richer query options available locally.

This architecture ensures that data remains at its source while enabling unified access across institutions. The ongoing technical development of the FCS within Text+ supports the integration of diverse data types into a coherent search infrastructure, promoting interoperability, scalability, and sustainable access to linguistic research data.

4.4. Data Processing and Analysis

Data processing is a central aspect for any infrastructure. Text+ provides a number of tools for the analysis of language-based data. Those tool predate Text+, but were developed by Text+ partners.

The WebLicht environment is a web-based tool for the automatic annotation of text corpora (Hinrichs et al., 2010). It allows users to define and execute workflows that make use of tokenizers, part of speech taggers, parsers, and other tools. Such tools must be available as web services and their input and output behaviour must be described in a formal format (TCF). This format allows WebLicht to orchestrate predefined or customized processing chains. The resulting annotations can then be visualized in an appropriate way, such as in a table or tree format. WebLicht provides access to such analysis tools without a user needing to install other software than their standard web browser.

The NLP pipeline *MONAPipe*[21] ("Modes of Narration and Attribution Pipeline") is jointly developed by Text+ partners under the coordination of SUB Göttingen (Dönicke et al., 2022; Barth et al., 2023). MONAPipe is an independent Python library built upon the spaCy framework (Honnibal et al., 2020).[22] The pipeline integrates dedicated classifiers from several disciplines – including Literary Studies, Digital Humanities, and (Computational) Linguistics – by means of spaCy's custom component functionality.[23] In this way, developments from the community become accessible to a wide range of users who can enrich texts with annotations from the pipeline. SUB ensures the maintenance and long-term sustainability of the pipeline, and required resources are made available via a Dataverse repository hosted on GRO.data (re3data.org, 2023), the research data platform operated by the GWDG.

4.5. Legal Issues: Derived Text Formats

The reuse and publication of textual data in research is often constrained by legal frameworks, particularly copyright law and data protection regulations. This is especially relevant for texts from the 20th and 21st centuries, where full-text access is frequently restricted. Within the TA Collections, the concept of *Derived Text Formats* (DTFs) provides a structured approach to address these challenges.

DTFs are systematically transformed versions of original texts, created through a combination of information enrichment, targeted reduction and change of order. The goal is to retain analytical value while ensuring that the resulting data no longer fall under copyright protection or enable reconstruction of the original text. This transforma-

[21] https://pypi.org/project/monapipe

[22] https://spacy.io

[23] https://textplus.pages.gwdg.de/collections/mona-pipe/getting_started/component_overview

tion process includes operations such as deletion, replacement, generalization, and randomization, applied at various levels of granularity.

By applying these transformations, DTFs make possible the publication of research data without infringing on the rights of authors or publishers of the original data. This is particularly important for open data infrastructures, where unrestricted access must be balanced with legal compliance. The national standard DIN 19461, currently under development, provides detailed guidelines for documenting the transformation process.

In the context of Text+, DTFs serve as a legal and technical bridge between legacy textual resources and FAIR data principles. They facilitate lawful reuse, support reproducible research, and promote interoperability across disciplines—while respecting the legal boundaries of the source material. For a conceptual and practical foundation of DTFs, see (Schöch et al., 2020). In Text+, a consortium-wide working group investigates how to best create and make available DTFs, fostering activities across interested partner organizations.

4.6. Data Depositing and Help Desk

Text+ also takes on research data from researchers located outside the Text+ partner organizations. Apart from the long-term archiving aspect, such research data becomes accessible to a wider audience under the Text+ umbrella. Interested researchers can contact Text+ via the helpdesk or the bi-weekly Research Rendezvous.[24] The helpdesk is a joint service provided by all task areas and their data centres. Both the general contact form[25] and the specific data depositing form[26] on the Text+ website automatically generate a ticket in the helpdesk system. There, the tickets are assigned to the appropriate TA. When a request for data depositing is received, a virtual meeting is arranged with the researchers to clarify what data is involved and in what formats it is available to determine the best centre to archive the data.

4.7. T+ Labs: GraphRAG-based Search

A large consortium such as Text+ is better suited to experiment with cutting-edge technology than an individual centre, whose foremost priority it to deliver robust services to its user base. One such technology are large language models, who have had a huge impact on many scientific disciplines, in addition to their relevance in industrial and real-world use cases. Such LLMs can also be put into use for improving the Findability aspect of FAIR.

To address the query formulation challenges inherent in heterogeneous research data catalogues searched by users from different disciplines and academic areas, a GraphRAG-based search routine has been implemented. By integrating structured metadata and unstructured contextual information from the Text+ Registry with graph-based semantic relationships, an LLM-assisted retrieval environment is created that goes beyond traditional keyword matching.

The core of this approach extends traditional retrieval methods through the integration of knowledge graphs and retrieval augmented generation (Lewis et al., 2020). This approach facilitates a context-aware access to the Text+ Registry by capturing both explicit metadata relationships and implicit semantic connections across diverse resource descriptions. The implementation follows a two-stage workflow. In the first stage, structured metadata fields containing largely unambiguous information (including PERSON, INSTITUTION, LANGUAGE, CORPUS, RESEARCH_METHOD, LINGUISTIC_FEATURE, and others) are mapped into the knowledge graph representation alongside a vector database. In the second stage, contextual information from full-text descriptions is processed and ingested into the knowledge graph using multilingual sentence transformers and generative LLMs. Custom academic prompts optimize the extraction process for scholarly terminology and research methodology descriptions common in language resources. This unstructured text enrichment operates without pre-specified entity constraints, allowing the system to discover semantic relationships automatically through the content itself. A preview is available with the Text+ Registry Searchlab[27].

5. Discussion and Conclusion

The hub approach of Text+ requires a division of labour between "Text+ Central" and its distributed partners. This division of labour can be observed along two main aspects.

Organizational-wise, the Text+ hub is a virtual centre; there is no Text+ central headquarter, neither significant designated headquarter staff, nor a central budget. In fact, most of the resources that are designated to central services stem from the partners. The organization of Text+ as a loosely-coupled federation of partner centres has a number of benefits. In particular, the expertise of the partners remains locally grounded, together with the communities that have been built and which they continue to serve. Also, working towards the Text+ umbrella has brought the partners closer together

[24] https://events.gwdg.de/category/208

[25] https://text-plus.org/en/helpdesk

[26] https://text-plus.org/en/daten-dienste/depositing

[27] https://fdm.awhamburg.de/registry-searchlab

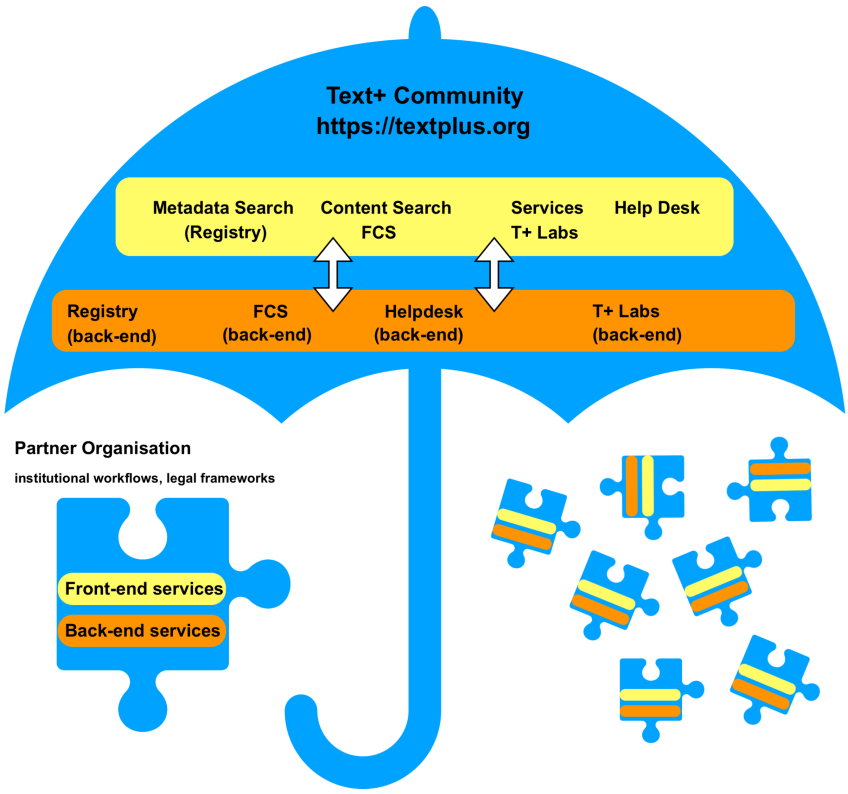

Figure 2: Text+ umbrella for central components in a federated structure.

as it has fostered many initiatives across them. It is worth to note that the division of labor between partners is defined in the Text+ work programme, which has a five year span.

From the technical perspective, the main mode of access to Text+ central services is the Text+ website. It is the main portal from which services such as the Registry, Federated Content Search, and Help Desk are made accessible to the interested public. These central services, however, are developed and hosted at different Text+ partner institutions, and they depend, in turn, on the front-end and back-end services of all partners, see Fig. 2.

The Text+ infrastructure depends on the partners making available and share their data, metadata, services, and expertise. Having their resources and services in the Text+ show display increases their visibility and reuse, and hence, contributes to a more FAIR research data management.

Text+ Governance Text+ with its 30+ partner institutions is governed by several key elements. The *Scientific Board* provides scientific leadership and makes strategic decisions on the development of the Text+ data, tools, and services. It is supported by the *Steering Committee*, which is responsible for the implementation of the work programme; this includes professional and financial oversight of ongoing activities. The *Institutional Board*, which is constituted by leaders of the (co)applying institutions, supports strategic and cross cutting matters across the consortium. Moreover, each task area (*i.e.*, data domain) in Text+ has a *Coordination Committee* that also has Text+ external experts as members. Their task is to continuously evaluate and expand the portfolio of data, tools, and services in each domain. In addition, Text+ is represented within the NFDI Association by an elected (deputy) spokesperson to align activities with other NFDI projects. Many Text+ partners are also involved in European infrastructures such as CLARIN and DARIAH; they are active scholars in their respective field of research, and hence part of the communities they serve. Community building is supported by frequent Text+ events.

In comparison to the pan-European CLARIN infrastructure, which is an independent legal entity (European Research Infrastructure Consortium, short ERIC), Text+ is not a separate legal entity, but simply a part of the NFDI. And while Text+ is organized by the three data domains Collections, Lexical Resources, Editions and its Infrastructure/Operations unit, CLARIN is organized in terms of its national nodes. Also note that the scientific board of Text+ is a decision-making body, while it only has an advisory role in CLARIN. Also, Text+ committees can exercise their advisory function with financial influence – so-called *flex funds* are decided with their involvement – while in CLARIN, committees have an advisory function only.

6. Ethics Statement

This paper does not involve experiments with human participants, nor does it include sensitive personal data or ethically problematic content. The linguistic resources discussed are available in the Text+ Registry and the Text+ FCS.

7. Acknowledgements

The work for this paper has been carried out within the Text+ National Research Data Infrastructure Consortium in Germany, funded by the German Research Foundation, grant number 460033370, and the contributions by the respective home institutions of the authors.

8. Bibliographical References

Florian Barth, Yannic Bracke, José Calvo Tello, George Dogaru, Tillmann Dönicke, Keli Du, Stefan E. Funk, Philippe Genet, Mathias Göbel, Lennart Keller, Daniel Kurzawe, Ubbo Veentjer, and Lukas Weimer. 2023. MONAPipe: Modular Natural Language Processing Pipeline for Digital Humanities.

Kristin Bührig. 2025. Curation Policy für das Forschungsdatenrepositorium der Universität Hamburg – Community des Hamburger Zentrums für Sprachkorpora (HZSK).

Kristin Bührig and Bernd Meyer. 2009. Dolmetschen im Krankenhaus (DiK).

Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP Architecture - Diving in the Deep Sea of Corpus Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3586–3591. Paris: European Language Resources Association (ELRA) 2016, Portoroz, Slovenia.

Nils Diewald and Eliza Margaretha. 2016. Krill: KorAP search and analysis engine. *Corpus Linguistic Software Tools. Journal for Language Technology and Computational Linguistics (JLCL)*, 31(1):73–90.

Tillmann Dönicke, Florian Barth, Hanna Varachkina, and Caroline Sporleder. 2022. MONAPipe: Modes of narration and attribution pipeline for German computational literary studies and language analysis in spaCy. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 8–15, Potsdam, Germany.

Christoph Draxler and Klaus Jänsch. 2004. SpeechRecorder - a universal platform independent multi-channel audio recording software. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 559–562, Lisbon, Portugal. European Language Resources Association (ELRA).

Christoph Draxler and Klaus Jänsch. 2008. WikiSpeech – A Content Management System for Speech Databases. In *Proc. Interspeech*, pages 1646–1649, Brisbane.

Christoph Draxler and Julian Pömp. 2022. OCTRA – An Innovative Approach to Orthographic Transcription. In *Proc. Interspeech 2022*, pages pages 5217–5218, Incheon, Korea.

Christoph Draxler and Julian Pömp. 2024. Octra Backend - eine skalierbare Infrastruktur für Transkriptionsprojekte. In *Proc. ESSV 2024*, Regensburg. TUDpress.

Tobias Gradl, Christoph Kudella, Harald Lordick, and Daniela Schulz. 2024. Towards a registry for digital resources – the text+ registry for editions. *Datenbank Spektrum*, 24(2):151–160.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – A lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Somerset, NJ. Association for Computational Linguistics.

Erhard Hinrichs and Thorsten Trippel. 2024. Text+ – Concept and Benefits for Empirical Researchers. *Cybernetics and Information Technologies*, 24(4):143 – 163.

Marie Hinrichs, Thomas Zastrow, and Erhard Hinrichs. 2010. Weblicht: Web-based lrt services in a distributed escience infrastructure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, page 489 — 493, Valletta, Malta.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength Natural Language Processing in Python.

ISO 24619. 2011. Language resource management – persistent identification and sustainable access (pisa). Standard ISO 24619:2011, International Organization for Standardization, Geneva, Switzerland.

ISO 24622-1. 2015. Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model. International Standard, International Organization for Standardization (ISO), Geneva.

ISO 24622-2. 2019. Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: Component metadata specification language. International Standard, International Organization for Standardization (ISO), Geneva.

ISO 24624. 2016. Language resource management — transcription of spoken language. Standard ISO 24624:2016, International Organization for Standardization, Geneva, Switzerland. Confirmed in 2022. This version remains current.

Markus Jochim. 2017. Extending the EMU Speech Database Management System: Cloud Hosting, Team Collaboration, Automatic Revision Control. In *Proceedings of Interspeech 2017, Stockholm, Sweden*, pages 813–814, Stockholm, Sweden.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Thomas Kisler, Florian Schiel, and Han Sloetjes. 2012. Signal Processing Via Web Services: The Use Case Webmaus. In *Proceedings Digital Humanities*, pages 30–34, Hamburg.

Erik Körner, Thomas Eckart, Uwe Kretschmer, Axel Herold, Frank Wiegand, Frank Michaelis, Matthias Bremm, Louis Cotgrove, Thorsten Trippel, Felix Rau, Anne Klee, Daniel A. Werning, Dominik Blöse, and Claus Zinn. 2025. *Federated Content Search for Lexical Resources (LexFCS): Specification*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Harald Lüngen and C. M. Sperberg-McQueen. 2012. A tei p5 document grammar for the ids text model. *Journal of the Text Encoding Initiative*, (3).

Scott Martens. 2013. TüNDRA: A Web Application for Treebank Search and Visualization. In *Proceedings of The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 133–144.

re3data.org. 2023. GRO.data. Last accessed: 2025-04-09.

Ahmed Saleh, Dirk von Suchodoletz, Ines Pisetta, Thorsten Trippel, Jan Leendertse, Jonathan Bauer, and Klaus Tochtermann. 2025. InvenioRDM in NFDI: Advancing FAIR Data Across Domains. In *Proceedings of the 2nd Conference on Research Data Infrastructure (CoRDI)*. Zenodo.

Thomas Schmidt. 2012. EXMARaLDA and the FOLK tools — two toolsets for transcribing and annotating spoken language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 236–240, Istanbul, Turkey. European Language Resources Association (ELRA).

Oliver Schonefeld, Thomas Eckart, Thomas Kisler, Christoph Draxler, Kai Zimmer, Miroslav Ďurčo, Yury Panchenko, Heike Hedeland, Anke Blessing, and Olga Shkaravska. 2014. Clarin federated content search (clarin-fcs) – core specification. Technical Report CE-2014-0316, CLARIN ERIC.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften (ZfdG)*, 5. Open Access.

Han Sloetjes, Albert Russel, and Alex Klassmann. 2007. ELAN: a free and open-source multimedia annotation tool. In *Proc. Interspeech*, pages 4015–4016, Antwerp.

TEI Consortium. 2025. *TEI: Guidelines for Electronic Text Encoding and Interchange*, p5 version 4.10.2 edition. Text Encoding Initiative Consortium. Last updated on 4th September 2025, revision bcfa98f42.

Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardelleni. 2010. Virtual language observatory: The portal to the language resources and technology universe. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pages 900–903. European Language Resources Association (ELRA).

Ubbo Veentjer, Stefan Buddenbohm, José Calvo Tello, Stefan E. Funk, Ralf Klammer, Nanette Rißler-Pipka, Alex Steckel, Lukas Weimer, George Dogaru, and Mathias Göbel. 2025. Fluffy import: Preserving humanities research data with the textgrid repository. *Transformations: A DARIAH Journal*.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Raphael Winkelmann, Jonathan Harrington, and Klaus Jänsch. 2017. Emu-SDMS: Advanced Speech Database Management and Analysis in R. *Computer Speech and Language*.

9. Language Resource References

Federated Content Search. Text+ – Consortium in the German National Researchdata Infrastructure (NFDI).

Registry - The Text+ Catalogue. Text+ – Consortium in the German National Researchdata Infrastructure (NFDI).