Text+

nfdi
Nationale
Forschungsdaten
Infrastruktur

utilize and preserve text- and
language-based research data





proposal for the second funding phase

full version

Cover page and subsequent illustrations:
Material Design icons used under license Apache 2.0

Abbreviations

For the names of the institutions that are part of the proposal, please refer to the tables in 1: General information.

AARC	Authentication and Authorisation for Research and Collaboration
AI	Artificial Intelligence
ALLEA	European Federation of Academies of Sciences and Humanities
API	Application Programming Interface
BARTOC	Basic Register of Thesauri, Ontologies & Classifications
BLV	Bund-Länder-Vereinbarung (federal-state agreement)
CARE	Collective Benefit, Authority to Control, Responsibility, Ethics
CENL	Conference of European National Librarians
CLARIN	Common Language Resources and Technology Infrastructure
CMDI	Component Metadata Infrastructure
CRC	Collaborative Research Centre (SFB, Sonderforschungsbereich)
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DCAT	Data Catalog Vocabulary
DevOps	Development and Operations
DGfS	Deutsche Gesellschaft für Sprachwissenschaft (linguistic society of Germany)
DHd	Digital Humanities im deutschsprachigen Raum (Digital Humanities in German-speaking countries)
DTF	Derived Text Formats (Abgeleitete Textformate)
DFG	Deutsche Forschungsgemeinschaft (German Research Foundation)
DIN	Deutsches Institut für Normung (German Institute for Standardisation)
DOI	Digital Object Identifier
EFNIL	European Federation of National Institutions for Language
EOSC	European Open Science Cloud
ERIC	European Research Infrastructure Consortium
FAIR	Findable, Accessible, Interoperable, Reusable
FCS	Federated Content Search
FID	Fachinformationsdienst (Specialised Information Service)
FITS	Flexible Image Transport System

GDPR	General Data Protection Regulation
GKFI	Geistes- und Kulturwissenschaftliche Forschungsinfrastrukturen e.V. (Association for research infrastructures in the humanities and cultural studies)
GND	Gemeinsame Normdatei (Integrated Authority File)
GPT	Generative pre-trained transformers
GUI	Graphical User Interface
HPC	High-Performance Computing
HTR	Handwritten Text Recognition
IFLA	International Federation of Library Associations and Institutions
ISO	International Organization for Standardization
ITIL	IT Infrastructure Library
IIIF	International Image Interoperability Framework
KPI	Key Performance Indicator
KV	Konsortialversammlung (Consortial Assembly)
LDS	European Language Data Space
LIBER	Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries
LIDO	Lightweight Information Describing Objects
LLM	Large Language Model
LOD	Linked Open Data
MARC21	Machine-Readable Cataloging
MCP	Model Context Protocol
METS	Metadata Encoding and Transmission Standard
ML	Machine Learning
MODS	Metadata Object Description Schema
NFDI	National Research Data Infrastructure
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OER	Open Educational Resources
OPERAS	Open Scholarly Communication in the European Research Area in the Social Sciences and Humanities
OWL	Web Ontology Language

PID	Persistent Identifier
PISA	Persistent Identification of Language Resources
POS	Part of Speech
PROV-O	PROV (=Provenance) Ontology
RAG	Retrieval Augmented Generation
RDF	Resource Description Framework
RDA	Research Data Alliance
RDM	Research Data Management
RDMO	Research Data Management Organiser
RfII	Rat für Informationsinfrastrukturen (German Council for Scientific Information Infrastructures)
SPA	Software Product Architecture
SPARQL	Query language for databases
SSH	Social Sciences and Humanities
SSHOC	Social Sciences and Humanities Open Cloud
TEI	Text Encoding Initiative
TF	Task Force
URI	Uniform Resource Identifier
URN	Uniform Resource Name
VIAF	Virtual International Authority File
VD17/VD18	Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. und 18. Jahrhunderts (Union Catalogue of Books Printed in German Speaking Countries in the 17th and 18th Century)
W3C	World Wide Web Consortium
WAPCAS	Web Annotation Protocol Compliant Annotation Store
WG	Working Group
XML	Extensible Markup Language

Contents

1. General information **7**
Name of the consortium in English and German 7
Summary of the proposal 7
Zusammenfassung 8
Applicant institution 8
Spokesperson 9
Co-applicant institutions 9
Co-spokespersons 9
Participating Institutions 10
DFG review boards that reflect the subject orientation of the proposed consortium 11
2. Scope and objectives **12**
2.1 Research domains or research methods addressed by the consortium, specific aim(s) 12
2.2 Objectives and measuring success 20
3. Consortium **22**
3.1 Composition of the consortium and its embedding in the community of interest 23
3.2 The consortium within the NFDI and the national academic research system 27
3.3 International networking 31
3.4 Organisational structure and viability 33
3.5 Operating model 36
4. Research data management strategy **38**
4.1 Scientific relevance and quality of the measures 39
4.2 Metadata standards 42
4.3 Implementation of the FAIR principles and data quality assurance 44
4.4 Services provided by the consortium 45
4.5 Impact of changes of external conditions / constraints 47
5. Work programme **49**
5.1 TA1: Library and Archive Holdings as Research Data 59
5.2 TA2: Language Corpora and Elicited Data 68
5.3 TA3: Lexical Resources 75
5.4 TA4: Editions 85
5.5 TA5: Infrastructure/Operations 95
5.6 TA6: Administration 106
6. Additional aspects **111**
6.1 Equity and diversity 111
6.2 Declaration of generative AI and AI-assisted technologies in the writing process 112
Appendix 1: Bibliography and list of references **113**
Appendix 4: Additional Comments on Data Sheet **126**
Appendix 5: Text+ supplementary data sheet for consortia of the NFDI **129**

1. General information

Name of the consortium in English and German

Text+: Language- and Text-Based Research Data Infrastructure

Text+: Sprach- und textbasierte Forschungsdateninfrastruktur

Summary of the proposal

Text+ is a research data infrastructure dedicated to supporting language- and text-based research across the humanities and social sciences. Building on achievements from its first funding phase, Text+ enters phase II with a sharpened strategy focused on sustainability, innovation, and inclusivity.

The consortium brings together 41 partner institutions with expertise in the management, preservation, and reuse of diverse research data types, including library holdings, corpora, lexical resources, and scholarly editions. Central to its mission is the Text+ Data Space, an interoperable and certified federated infrastructure for storing, discovering, and processing language- and text-based data. All services are grounded in FAIR principles and aligned with national and European frameworks such as CLARIN, DARIAH, and EOSC.

Phase II responds to a transformed research landscape shaped by AI, large language models (LLMs), and the growing importance of language as a universal data interface. To address these shifts, Text+ defines four strategic themes: (1) language/text as a dominant interface to data; (2) the dynamic interplay between structured and unstructured data; (3) a human-centred research culture that embeds ethics and sustainability; and (4) the long-term sustainability of infrastructure and services.

These themes are operationalized in four data domain-specific task areas (TAs): 1) Library Holdings as Research Data, 2) Language Corpora and Elicited Data, 3) Lexical Resource, and 4) Editions, and two transversal task areas: 5) Infrastructure/Operation and 6) Administration. Overarching objectives include expanding community integration, enhancing AI capabilities, stabilizing services, FAIRifying legacy data, and deepening national and international collaboration. A strong emphasis is placed on training, outreach, and embedding RDM into academic curricula.

Text+ also plays a central role in the National Research Data Infrastructure (NFDI), contributing to cross-consortial services such as PID4NFDI, IAM4NFDI, and Jupyter4NFDI, and participating in the Base4NFDI initiative and NFDI governance bodies. Internationally, the consortium ensures interoperability with European infrastructures, engaging deeply with CLARIN, DARIAH, and OPERAS.

Through this work, it helps realize the vision of “data as a common good,” fostering innovation and participation in a digitally transformed academic world.

Zusammenfassung

Text+ ist eine Forschungsdateninfrastruktur, die sprach- und textbasierte Forschung in den Geistes- und Sozialwissenschaften unterstützt. Aufbauend auf den Erfolgen der ersten Förderphase, startet Text+ in Phase II mit einer geschärften Strategie, die auf Nachhaltigkeit, Innovation und Inklusivität fokussiert.

Das Konsortium vereint 41 Partnereinrichtungen mit Expertise in der Verwaltung, Bewahrung und Wiederverwendung diverser Forschungsdatentypen, darunter Bibliotheksbestände, Korpora, lexikalische Ressourcen und Editionen. Zentral ist der Text+ Data Space, eine interoperable, zertifizierte föderierte Infrastruktur zur Speicherung, Auffindung und Verarbeitung sprach- und textbasierter Daten. Alle Dienste basieren auf den FAIR-Prinzipien und sind auf nationale und europäische Rahmenwerke wie CLARIN, DARIAH und EOSC abgestimmt.

Phase II reagiert auf eine transformierte Forschungslandschaft durch KI, große Sprachmodelle (LLMs) und die wachsende Bedeutung von Sprache als universelle Datenschnittstelle. Text+ definiert hierfür vier strategische Themen: (1) Sprache/Text als dominante Datenschnittstelle; (2) das dynamische Zusammenspiel von strukturierten und unstrukturierten Daten; (3) eine menschenzentrierte Forschungskultur mit Ethik und Nachhaltigkeit; und (4) die langfristige Nachhaltigkeit von Infrastruktur und Diensten.

Diese Themen werden in vier datendomänenspezifischen Aufgabenbereichen – 1) Bibliotheksbestände als Forschungsdaten, 2) Sprachkorpora und Elicited Data, 3) Lexikalische Ressourcen und 4) Editionen – und zwei Querschnittsaufgabenbereichen – 5) Infrastruktur/Betrieb und 6) Administration – operationalisiert. Übergeordnete Ziele sind die Ausweitung der Community-Integration, Verbesserung der KI-Fähigkeiten, Stabilisierung der Dienste, FAIRifizierung von Altdaten und Vertiefung der nationalen und internationalen Zusammenarbeit. Besonderer Wert wird auf Schulung, Outreach und die Verankerung von Forschungsdatenmanagement (FDM) in akademischen Lehrplänen gelegt.

Text+ spielt auch eine zentrale Rolle in der Nationalen Forschungsdateninfrastruktur (NFDI), indem es zu konsortienübergreifenden Diensten beiträgt und aktiv an der Base4NFDI-Initiative sowie NFDI-Steuerungsgremien teilnimmt. International sichert das Konsortium die Interoperabilität mit europäischen Infrastrukturen, eng verbunden mit CLARIN, DARIAH und OPERAS.

Durch diese Arbeit hilft Text+, die Vision von „Daten als Gemeingut“ zu verwirklichen und Innovation und Partizipation in einer digital transformierten akademischen Welt zu fördern.

Applicant institution

Applicant institution	Location	Short
Georg-August-Universität Göttingen	Göttingen	UGOE/SUB

Spokesperson

Spokesperson	Institution, location	Short
Prof. Dr. Philipp Wieder	Georg-August-Universität Göttingen, Göttingen Göttingen State and University Library	UGOE/SUB

Co-applicant institutions

Co-applicant institutions	Location	Short
Berlin-Brandenburg Academy of Sciences and Humanities	Berlin	BBAW
German National Library	Frankfurt am Main	DNB
Humboldt-Universität zu Berlin	Berlin	HUB
Leibniz-Institute for the German Language	Mannheim	IDS
Marburg University/Research Center Deutscher Sprachatlas	Marburg	UMR
North Rhine-Westphalian Academy of Sciences, Humanities and the Arts	Düsseldorf	AWK-NRW
Saarland University	Saarbrücken	UdS
Technical University of Darmstadt	Darmstadt	TUDa
University of Tübingen	Tübingen	EKUT
Verbundzentrale (VZG) des Gemeinsamen Bibliotheksverbundes (GBV)	Göttingen	VZG

Co-spokespersons

Co-spokespersons	Institution, location	Task area(s)
Prof. Dr. Hanna Fischer	Marburg University/Research Center Deutscher Sprachatlas, Marburg	Lexical Resources
Philippe Genêt	German National Library, Frankfurt am Main	Library Holdings as Research Data
PD Dr. Alexander Geyken	Berlin-Brandenburg Academy of Sciences and Humanities, Berlin	Lexical Resources
Prof. Vivien Petras, PhD	Humboldt-Universität zu Berlin, Berlin	Infrastructure/Operations
Prof. Dr. Andrea Rapp	Technical University of Darmstadt, Darmstadt	Editions
Prof. Dr. Dr. h.c. Andreas Speer	North Rhine-Westphalian Academy of Sciences, Humanities and the Arts, Düsseldorf	Editions
Regine Stein	Verbundzentrale (VZG) des Gemeinsamen Bibliotheksverbundes (GBV), Göttingen	Infrastructure/Operations
Prof. Dr. Elke Teich	Saarland University, Saarbrücken	Library Holdings as Research Data
Prof. Dr. Andreas Witt	Leibniz-Institute for the German Language, Mannheim	Language Corpora and Elicited Data; Administration
Prof. Dr. Angelika Zirker	University of Tübingen, Tübingen	Language Corpora and Elicited Data

Participating Institutions

Table 1.1 Institutional contributions to Text+ (P = Participant)

#	Participating Institutions	Short	Location	Principal Investigators
P-01	Academy of Sciences and Humanities in Hamburg	AdWHH	Hamburg	Claudia Raap
P-02	Academy of Sciences and Literature | Mainz	AWLM	Mainz	Prof. Dr. Roland Kehrein
P-03	Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)	GWDG	Göttingen	Dr. Sven Bingert
P-04	Bavarian Academy of Sciences and Humanities	BAdW	München	Dr. Eckhart Arnold
P-05	The University of Wuppertal	BUW	Wuppertal	Prof. Dr. Patrick Sahle
P-06	Bielefeld University	UniBI	Bielefeld	Prof. Dr. Berenike Herrmann
P-07	Deutsche Schillergesellschaft e.V. Deutsches Literaturarchiv Marbach	DLA	Marbach am Neckar	Prof. Dr. Sandra Richter
P-08	Dresden University of Technology	TUDD	Dresden	Dr. Ralph Müller-Pfefferkorn
P-09	Forschungszentrum Jülich GmbH	JSC	Jülich	Björn Hagemeier
P-10	Freie Universität Berlin	FU	Berlin	Dr. Cord Pagenstecher
P-11	German Institute for Adult Education - Leibniz Centre for Lifelong Learning	DIE	Bonn	Dr. Kerstin Hoenig
P-12	Göttingen Academy of Sciences and Humanities in Lower Saxony	NAWG	Göttingen	Dr. Jörg Wettlaufer
P-13	Heidelberg Academy of Sciences and Humanities	HAdW	Heidelberg	Dr. Dieta Svoboda-Baas
P-14	Herzog August Bibliothek, Wolfenbüttel	HAB	Wolfenbüttel	Dr. Johannes Mangei
P-15	Hochschule Wismar, University of Applied Sciences, Technology, Business and Design	HSW	Wismar	Prof. Dr. Frank Krüger
P-16	Julius-Maximilians-Universität Würzburg	JMU	Würzburg	Dr. Torsten Roeder
P-17	Klassik Stiftung Weimar	KSW	Weimar	Dr. Gerrit Brüning
P-18	Ludwig-Maximilians-University Munich	LMU	München	PD Dr. Christoph Draxler
P-19	Max Weber Foundation - German Humanities Institutes Abroad	MWS	Bonn	Dr. Michael Kaiser
P-20	Paderborn University	UniPB	Paderborn	Prof. Dr. Andreas Münzmay

P-21	Ruhr University Bochum	RUB	Bochum	Jun.-Prof. Dr. Tatjana Scheffler
P-22	Salomon Ludwig Steinheim Institute of German-Jewish History (e.V.)	STI	Essen	Prof. Dr. Lucia Raspe
P-23	Saxon Academy of Sciences and Humanities in Leipzig	SAW	Leipzig	Uwe Kretschmer
P-24	Trier University	UniTR	Trier	Prof. Dr. Christof Schöch
P-25	University of Applied Sciences Darmstadt	h_da	Darmstadt	Prof. Dr. Stefan Schmunk
P-26	University of Bamberg	UniBa	Bamberg	Prof. Dr. Andreas Henrich
P-27	University of Cologne	UzK	Köln	Prof. Dr. Nils Reiter
P-28	University of Hamburg	UHH	Hamburg	Prof. Dr. Kristin Bührig
P-29	University of Freiburg	UFR	Freiburg	Prof. Dr. Achim Rabus
P-30	University of Rostock	UniRo	Rostock	JProf. Dr. Ulrike Henny-Krahmer

DFG review boards that reflect the subject orientation of the proposed consortium

1.11-02 Ancient Cultures (partly); Greek and Latin Philology;

1.14 Linguistics

1.15 Literary Studies

1.16 Social and Cultural Anthropology; Non-European Cultures; Jewish Studies and Religious Studies (partly)

1.18 Philosophy

1.25 Jurisprudence (partly)

For details, please refer to appendix 5 (fields 1022 and 1023).

2. Scope and objectives

2.1 Research domains or research methods addressed by the consortium, specific aim(s)

2.1.1 Text+ after the first five years

Text+ is designed to support a broad spectrum of research domains and methods within the language- and text-based humanities as well as social sciences (Hinrichs et al., 2022). These include, among others, philology and literary studies, philosophy, linguistics, translation studies, digital humanities, cultural studies, also their specialized subdisciplines such as computational and applied linguistics. Among the disciplines are also the social sciences that are analysing language and texts for their research. The consortium supports these domains through a combination of data services, infrastructure, consulting, and training (see 3.1.3), all grounded in the FAIR principles (Wilkinson et al., 2016) and, where applicable, also informed by work based on the CARE principles (Carroll et al., 2020). The Text+ Data Space, comprising certified Text+ Centers (e.g. nestor, DIN), the Text+ Registry, and the Federated Content Search (FCS), has become a cornerstone for interoperable and reusable research data (see also 4.3).

In the context of Text+, text refers to language data in its concrete, observable form, including not only written or printed language, but also spoken, transcribed, encoded, and digitally structured manifestations. *Text* is thus understood as a material instantiation of language, shaped by linguistic, medial, cultural, and technological factors, and made accessible for scholarly inquiry through metadata, annotations, and infrastructural embedding. Text+ specifically focuses on language data at all scales, including large-scale corpora and small language models.

The research methods supported by Text+ span corpus compilation and annotation, lexicographic modelling and querying, scholarly editing and digital edition curation, metadata standardization (see 4.2) and authority file integration (see section 4, in particular 4.2 and 4.3), natural language processing and text mining, FAIRification and data depositing workflows (see 4.3). These methods have been implemented in the first funding phase across a wide range of data types, including multimodal and spoken language corpora, historical and contemporary texts, as well as lexical resources. The consortium has developed and maintained tools and services that enable researchers to work with these data types in a sustainable and interoperable manner (see 4.4).

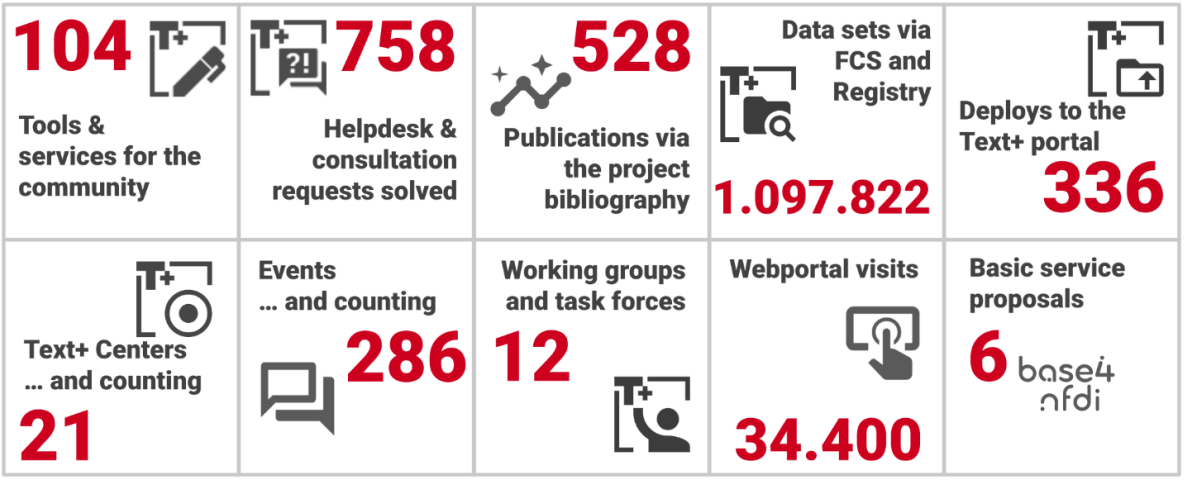


Figure 2.1: Added value for the community in phase I [data for webportal: 2023–25]

Text+ has had a significant impact on the research landscape. With more than 30 institutions involved, including those added during the first phase, the consortium draws on a rich foundation of over 1.097.822 data sets curated and maintained by its centers and partner organizations. These data sets–many of which originate from earlier projects and institutional repositories–are being successively reviewed, harmonized, and integrated into the Text+ Data Space. Internationally, Text+ has established strong collaborations with CLARIN (Fišer & Witt, 2022; Branco et al., 2023), DARIAH (Hedges et al., 2013), SSHOC (SSHOC Consortium, n.d.), and the European Open Science Cloud (EOSC) (EOSC Association, n.d.), aligning its services with European standards and ensuring global accessibility (see 3.3). The consortium has also taken a leadership role in cross-consortial initiatives, contributing to Base4NFDI (Bernard et al., 2023) and participating in sections and working groups within the NFDI governance structure (see 3.2.3).

Training and outreach have been central to the consortium's mission (see 3.1.3), exemplified by workshops, helpdesk consultations, and community events such as the Research Rendezvous, plenaries, and community-oriented workshops, overall promoting a culture of openness, reuse, and methodological innovation.

Text+ has also responded strategically to emerging challenges, including the disruptive developments in AI driven by the evolution of generative models (see 4.5). Through these endeavours, Text+ has positioned itself as a central infrastructure for language- and text-based research in Germany and beyond, addressing fragmentation in data practices and fostering sustainable research practices.

2.1.2 Grand challenges as catalysts: building the Text+ infrastructure

In phase I of Text+, the conceptual and structural development of the consortium was guided by three grand challenges shaping work programme, architecture, and services:

- **Grand Challenge I:** Mastering the diversity of research data and the diversity of communities of interest for language and text
- **Grand Challenge II:** Assisting humanities scholars in the re-use, production, preservation, and innovative use of research data
- **Grand Challenge III:** Developing a governance model that supports joint responsibility and consensus building among infrastructure providers and infrastructure users as well as collaboration across disciplinary boundaries

Text+ has supported a broad spectrum of research domains, particularly within the language- and text-oriented humanities and the social sciences. In phase I, the consortium's work plan covered three data domains, which were also task areas in the organizational structure: "Collections," focusing on language annotation, corpus compilation, and querying; "Lexical Resources," dedicated to the development of lexical data; and "Editions," centered on scholarly editing. All activities are grounded in FAIR data principles.

Under **Grand challenge I**, Text+ successfully designed an infrastructure capable of accommodating the full range of diverse research data types, formats, and disciplinary requirements. The creation of the Text+ Data Space – with its distributed architecture of certified centers, modular services, and flexible workflows – has proven that even highly heterogeneous data can be made FAIR across contexts. The establishment of the data domains was informed by expertise in the respective domains, while the utilisation of the same structures in the form of measures ensured cohesion and interoperability.

Grand challenge II has been operationalised through an ambitious portfolio of support and integration activities. The helpdesk, the consulting framework, regular community workshops, and the cooperation projects have enabled researchers to not only reuse and preserve their data but to also participate in shaping the infrastructure. The Text+ Registry, FCS interfaces, depositing services, the use of authority file information (GND) (German National Library, n.d.a) and training materials all exemplify this commitment to empowerment and scholarly inclusion.

Grand challenge III has been addressed through a multi-layered governance model that ensures both operational efficiency and community participation. Coordination Committees (CCs) representing the scholarly communities have played a central role in advising, reviewing, and guiding consortium decisions. A culture of joint responsibility has emerged, supported by legal frameworks, transparent decision-making processes, and regular community engagement formats such as the Text+ plenary and Research Rendezvous. This has also allowed Text+ to support disciplinary diversity, reflect community needs, and adapt to structural changes within the NFDI landscape.

The work on these grand challenges has laid the foundation for an infrastructure that is not only technically sound and scalable but also deeply embedded in the research cultures it serves. As a result, the original grand challenges have become operational competencies that are integrated into the daily routines and strategic planning of the consortium.

2.1.3 Overarching topics of phase II: responding to a changing research landscape

Since the inception of Text+, the research and infrastructure landscape has undergone a profound transformation. The emergence of large language models (LLMs) (Minaee et al., 2025) and AI-driven methods, the expansion of language as a universal interface across disciplines and media, and the growing urgency of ethical, societal, and sustainability-related questions have introduced new complexities that existing infrastructures must address. These shifts mark not a rupture, but an intensification of the original challenges, requiring refined strategies and conceptual perspectives for phase II.

In response, the consortium has expanded the scope of its research domains and methods to include, e.g., research on under-resourced and endangered languages, multimodal and machine-generated textual corpora, and computational approaches to lexical and semantic annotation and linking. This evolution is reflected in the revised work programme for phase II (see section 5, in particular 5.0) as follows:

- Introduction of new task areas, enabled by the project's built-in adaptability:
  - *Library holdings as Research Data* (see 5.1), previously part of the task area Collections in a separate thematic cluster "unstructured text", focusing on the structured reuse of digitized and born-digital collections available in library holdings. The specific requirements for rights management, interoperability, and AI-based processing set these data apart from other domains.
  - *Language Corpora and Elicited Data* (see 5.2), previously part of the task area Collections in the thematic clusters "contemporary language" and "historical texts", focuses on all kinds of corpora, from linguistics to literary studies and beyond, also including elicited data resulting from experiments, surveys, etc. Given the diversity of data sources and the increasing integration of AI workflows, a dedicated task area is deemed necessary.
- Refinement of existing task areas – *Lexical Resources* (see 5.3), *Editions* (see 5.4), *Infrastructure and Operations* (see 5.5) – aligning them with current developments regarding AI, ethics, and interoperability with relevance for Text+/the NFDI.

To respond to the changing research landscape and refine the specific aims of the consortium, Text+ introduces four strategic themes reflecting the aims that will guide the consortium's conceptual and infrastructural development in phase II: (1) Language/text as dominant interface to data; (2) Dynamic interplay between structured and unstructured

data; (3) Human-centered research culture; (4) Longevity of the Text+ research data infrastructure.

Strategic Theme 1: Language/text as dominant interface to data: Text is a carrier of content and a central interface for the representation, analysis, and transformation of data across disciplines. With the recent advances in AI and LLMs, awareness has spread across all disciplines that text is not merely a medium but also a method for describing, interpreting, generating, and structuring knowledge. Whether in automatically describing or generating images based on textual input, transcribing speech, generating code from language prompts, or constructing multimodal representations, language and text are at the heart of digital research processes. The aim is to open up new pathways for engaging with research data in the humanities and social sciences, particularly in areas such as text classification, annotation, and application programming. Crucially, the integration of LLMs into these processes benefits significantly from domain-specific language expertise and curated textual resources - areas where Text+ offers unique strengths. At the same time, the consortium is committed to critically assessing the implications of LLMs for scientific research (Triantaphyllou & Polatidis, 2023), ensuring that innovation is accompanied by reflection, thus contributing both infrastructure and expertise to the evolving landscape of AI-driven research.

Strategic Theme 2: Dynamic interplay between structured and unstructured data aims at the interplay between “structured” and “unstructured” data across the data domains of language- and text-based research as well as between different types of structure emerging from specific scientific inquiries. Here, “structured” and “unstructured” are seen as poles on a cline where data is never completely structured or unstructured. If the identified structures are commensurate with one field of research, they may be regarded as unstructured in a different one. What is more, for a given object or data set, there are typically multiple structures (e.g. document structure, tokens, annotations, etc.) that are (more or less) interconnected. Strategic theme 2 addresses the specific character of language- and text-based data from more or less structured to multiple, interrelated structure(s). It is vital to acknowledge that the inherent multiplicity of perspectives on text and language data in the humanities presents several challenges from the perspective of a research data infrastructure: Which structures are foregrounded, which are backgrounded, or even omitted entirely? How do researchers negotiate competing structural logics? What implications arise when infrastructures privilege one structuring mode over another?

The process of making structures explicit transforms implicit knowledge into documented information, thereby enriching the data and enhancing its quality. This, in turn, significantly broadens its potential for reuse. Hence, by enriching data with layered, interoperable structures, Text+ improves data quality, enables more nuanced analysis, and opens up new

possibilities for reuse across disciplines, institutions, and research questions. Strategic theme 2 aims at models that accommodate layered, partial and coexisting structures, allowing researchers to navigate, compare, and selectively engage with structural multiplicity without being forced into reductive frameworks. In addition, cutting across data domains, we observe and reflect upon how researchers move between unstructured and structured representations. This is reflected also by the following strategic theme.

Strategic Theme 3: Human-centered research culture places the human at the center of digital transformation: It aims at (re-)anchoring the human dimension in an increasingly automated, AI-driven world of text and language processing by addressing cultural and societal values as well as issues of data sovereignty. Text+ stands for the ethical, just, and culturally reflective use of text- and other language-based data and AI technologies. Beyond mitigating algorithmic bias and complying with legal frameworks, this challenge raises fundamental questions about how technological developments can be framed, critically examined, and meaningfully shaped. It is at the heart of the twin transformation, i.e. working on digitisation while at the same time increasing sustainability across the realms of society. Moreover, the use of AI technologies–especially large language models (LLMs)–requires substantial energy resources. Text+ is committed to a critical assessment of both the benefits and drawbacks of these technologies, with a particular focus on their ecological footprint. This strategic theme contributes to a sustainable, human-centred research culture in which technology serves the human–not the other way around. It seeks to embed the interpretive depth and ethical awareness of the humanities in the heart of digital research infrastructures. This integration includes ensuring that infrastructures remain inclusive, open to plural perspectives, and capable of supporting critical, contextualized research (see 6).

Strategic Theme 4: Longevity of the Text+ research data infrastructure: A core aim of the NFDI as a whole is to establish sustainable services, processes, and structures that will transform into a reliable research data management infrastructure for all academic disciplines in Germany. There is a common challenge for all of the consortia within the NFDI, namely creating something which is often called OneNFDI (Wissenschaftsrat, 2025) (see 3.2 and 3.5). To strategically contribute to this national endeavour and, at the same time, ensure that Text+ with its data, services, and expertise remains a reliable partner for its community of interest in the future, it is essential to strategically focus on this particular aspect of long-term sustainability in the second funding phase.

The foundations for longevity of the Text+ infrastructure have already been laid in the first funding phase (and actually long before, see 3.1). Central components, which are core to the realisation of this strategic theme, are the Text+ Data Space, the Text+ Centers, the Text+ Portal, the Text+ Service Catalogue, and the Text+ Helpdesk. Text+ will strategically

invest substantial resources within the second funding phase, complemented by no less extensive in-kind resources, to integrate these components into a sustainable whole, which includes required improvements to the Registry and FCS reported by the user community. Equally covered by this strategic theme is the co-ordinated evaluation and integration of Base4NFDI services (see 3.2.3).

Core to the realization of this strategic theme are the measures on “Federated Research Data Infrastructure” (TA1-5 M1; see 5.0), aiming at long-term FAIR data provision combined with sustainable operation of the Text+ Data Space (see also 3.5). These measures are complemented by the measure “Coherence” in the task area “Infrastructure/Operations” where policies are developed, interfaces to NFDI, EOSC, and other relevant infrastructures are maintained, and helpdesk processes are consolidated and streamlined.

Importantly, the strategic themes for phase II do not replace the original grand challenges of phase I but rather build on their foundation. The new strategic themes build upon, deepen, and recontextualize the achievements of phase I. They represent adaptive responses to ongoing technological, disciplinary, and societal transformations by addressing challenges and opportunities introduced by AI and large-scale language technologies, emphasizing data diversity and structural complexity, and ensuring that the infrastructures built remain inclusive, participatory, and ethically grounded.

The four strategic themes drive the continued development of the Text+ infrastructure which align with the goals of NFDI and the overarching objectives of federal and state research policy within NFDI.

In accordance with the Text+ vision to support the text- and language-oriented humanities and social sciences in leveraging digitization across research, teaching, and transfer, establishing a common data culture, the strategic themes address key transformations in how language and text function in research. They reflect the shift toward a common data culture grounded in FAIR text- and language-based data management.

In line with the vision of “Data as a common good” (NFDI e.V., n.d.d) and the federal-state agreement goals (e.g. interoperable infrastructures, discipline-accepted standards, reuse of existing data), the themes also:

- Foster collaboration across consortia and communities through a focus on shared infrastructures and services (NFDI Strategy 1, 2, 3) (NFDI e.V., n.d.d);
- Support internationalisation by aligning with European developments in other relevant infrastructures and Open Science (NFDI Strategy 4, EOSC alignment);
- Promote data competence and education by embedding digital literacy, ethics, and humanities methods into future infrastructure development (NFDI Strategy 5);
- Strengthen sustainability by guiding infrastructure development through strategically anchored, community-relevant challenges (NFDI Strategy 6).

With the four strategic themes for phase II, Text+ positions itself as a core actor in shaping a digital research infrastructure that is human-driven, ethically grounded, and technologically cutting-edge–and in so doing, fulfils its role within the NFDI ecosystem. Text+ will thus participate in the forthcoming process to ensure the long-term viability of the NFDI: reflecting the recent recommendations of the German Science and Humanities Council (Wissenschaftsrat, 2025) and the German Council for Scientific Information Infrastructures (Rat für Informationsinfrastrukturen, 2025), a transformation process aiming at permanently incorporating the NFDI into the German research system is expected to start within the second funding phase. Text+ will actively engage in further shaping the OneNFDI in close cooperation with all consortia and together with all stakeholders.

2.1.4 Project design: from planning to implementation

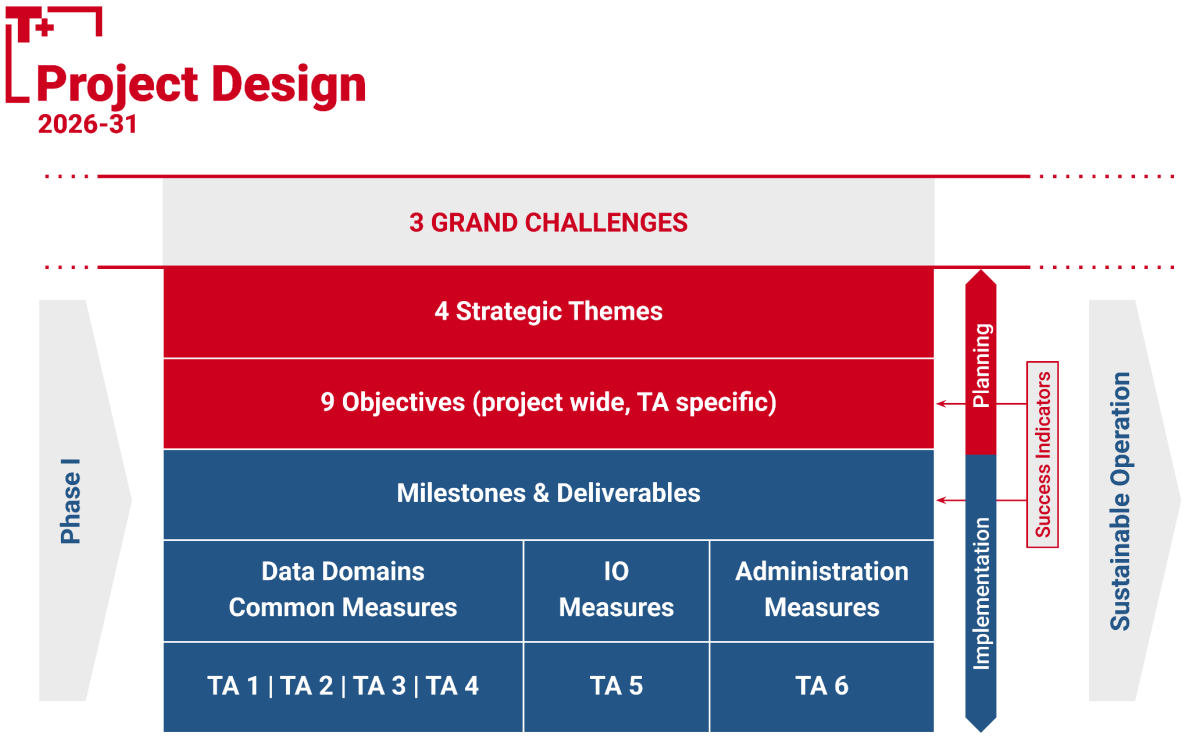


Figure 2.2: Project design connecting grand challenges, strategic themes, and task areas

With the grand challenges functioning as operational competencies, the four strategic themes form the foundation of Text+'s strategy for the upcoming funding phase. They guide the planning of phase II (see figure 2.2) and define the project's mid- to long-term strategic direction. These strategic themes translate into nine measurable, project-wide objectives (see 2.2) and span the full duration of phase II. Within the task areas, these objectives are further detailed into TA-specific objectives. These are typically more focused, short- to mid-term in duration (1–3 years), and are implemented through the respective measures and activities of the task areas (the latter not depicted for the sake of simplicity).

The milestones and deliverables define key development checkpoints and tangible outputs for Text+ to monitor and structure progress across the task areas and their measures. Success indicators are tied to the objectives and used to assess progress via milestones and deliverables. They connect planning and implementation.



The core instruments of the implementation are measures (mid- to long-term) and activities (short- to mid-term) with the data domains sharing the same common measures across all task areas to strengthen coordination and collaboration (see 5.0.1). The task areas themselves organise the consortium's work into six operational areas according to the work programme: four data domains, Infrastructure/Operations, and Administration (no dedicated subpage) (see 5.0).

In phase I, flexible funds were instrumental in supporting community growth and the integration of additional data sources. Going forward, this mechanism will be adapted to address dynamic, evolving needs and realise additional activities along the lines of the strategic themes and the project's objectives. These activities will, as those already planned within the task areas, target short- to mid-term objectives and act in response to community demand, emerging requirements, strategic opportunities, technological advancements, or the evolving national or European landscapes. Thus the consortium remains responsive, inclusive, and aligned with the evolving priorities within the NFDI. The usage of the flexible funds will be governed by TA6 Administration through the dedicated measure AD-M2 Flexible Funds (see 5.6). The respective processes have been designed and implemented already in phase I and will be adapted in phase II, respectively. Text+ has reserved around 10% of its overall funding for flexible funds (see 7.6).

2.2 Objectives and measuring success

In phase I, the consortium established itself as an infrastructure especially for the language- and text-oriented humanities (Text+ Consortium, 2024a). It addressed fragmentation in data practices, supported methodological diversity, and responded to the needs of its communities of interest through transparent, participatory governance. These objectives remain essential and will be maintained as structural tasks of the consortium.

In phase II, Text+ will shift its emphasis from initial provision and onboarding to strategic consolidation, conceptual refinement, and technical evolution. It will expand its reach to new communities, data types, and technologies, particularly in light of the growing role of AI, the increasing centrality of text as a data interface, and the pressing concerns relating to ethics and sustainability that digital research infrastructures are facing (see 5.0).

Accordingly, Text+ defines the following nine project-wide, strategic objectives for the second funding period. They serve as the foundation for the work programme and guide all measures and activities across the six task areas. In section 5, the various task areas are introduced, including their own TA-specific objectives and success indicators tailored to their respective scopes.

Each project-wide objective supports the specific aim of maintaining Text+ as a flexible, ethically grounded, and technologically robust infrastructure, and contributes to the overarching aims of the NFDI.

Objective 1: Further strengthen research community integration. Text+ will expand to new communities of interest, especially those not yet deeply embedded in infrastructure use (e.g. (oral) history, under-resourced languages, non-standardized language varieties, law, medicine, sociology, economics). Emphasis will be placed on onboarding, training, and participatory formats for scholars in underrepresented subfields, disciplines, and institutions.

Objective 2: Stabilise core Text+ services. Text+ will ensure the long-term availability, robustness, and usability of services such as the Text+ Registry, FCS, and depositing platforms. This includes technical scaling, security, service integration, and coordinated hosting strategies.

Objective 3: Enhance models for complex, interoperable data. Text+ will improve support for working with coexisting, layered, and evolving data structures, in line with the strategic theme 2. Models and services will be evaluated in terms of their flexibility, compatibility, and epistemological transparency.

Objective 4: Promote responsible use of AI, esp. LLMs, for language and text data. In response to recent shifts in the text-centered research landscape, Text+ will explore the integration of LLMs and related technologies for the use with language and text data, while critically examining their epistemic assumptions, risks, and infrastructural implications.

Objective 5: FAIRify new and legacy resources. Text+ will extend its FAIRification strategy to include digitized, born-digital, and legacy resources, working with project partners to ensure semantic indexing, persistent identifiers, and reusability.

Objective 6: Expand training, outreach, and user support. Text+ will continue to build the digital literacy of its user base through open materials, summer schools, tutorials, and direct support via the Helpdesk. It will tailor formats to different user profiles and needs, including early-career researchers.

Objective 7: Deepen engagement in the NFDI ecosystem. Text+ will remain an active contributor to cross-consortial services, metadata alignment, and standards development. It will work closely with Base4NFDI and participate in the NFDI sections.

Objective 8: Enhance international integration, particularly with EOSC and ERICs. Text+ will increase alignment with CLARIN, DARIAH, and other relevant platforms and contribute to the development of the NFDI EOSC National Node.

Objective 9: Secure long-term sustainability and governance. Text+ will define reliable operational, financial, and legal models for long-term service provision. This includes preparing for NFDI's future sustainability framework.

Table 2.1 Success indicators for the objectives

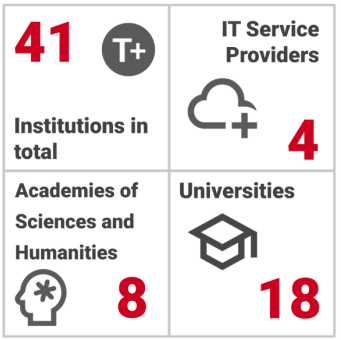
Objective	Success Indicator	Reference to Deliverable
1	Participation in Text+ training events; Number of Helpdesk tickets and consultations; Disciplinary breadth of disciplines represented in the consortium; Number of cooperations outside of the consortium; Integration of new actors; Number of users	TA 1-4, D0.5.2; TA6, D6.4.3 and appendix 5, fields 1021–1023, 1013–1017; TA 1-4, D0.1.5
2	Increase of services and maturity level; System stability; Service documentation	TA6, D6.4.3 and appendix 5, fields 4001–4008; TA5 M3, D5.3.1; TA5 D5.1.1, D5.1.2
3	Adoption of models; Metadata availability	TA 1-4 D0.4.1–6; TA 1-4 D0.1.3
4	Published guidelines, technical documentation, workshops; Integration of additional language models	TA 1-4 D0.3.1–7; TA 1-4 D0.3.2
5	Increase of number of data sets with FAIR-compliant metadata (PID+Metadata+Data exist for each data set); Inclusion in search services	TA 1-4 Milestone 0.1.3; TA5 D5.1.1, D5.1.2
6	Number of events and participants; User feedback	TA 1-4 D0.5.2; TA 1-4 D05.5
7	Adoption of services of Base4NFDI; Adoption of metadata NFDI recommendations; Section contributions; Alignment with NFDI-wide strategy documents	TA5 D5.3.4; TA1-4 D0.1.3; appendix 5, fields 1011–1012; TA6 D6.6.2; TA1-4 D0.1.1
8	International usage	TA 6 D6.3.6 and appendix 5, field 1014
9	Documented architecture; Procedures for each governance body; Operational model	TA 5 D5.4.4; TA6 D6.3.4 and D6.5.2; Together with all NFDI consortia

The success indicators will be key to select technical and operational KPIs in the work programme, which Text+ will use to evaluate its success through participatory feedback, regular consultation with its Coordination Committees, and strategic reviews in alignment with its mission. These objectives and indicators will be revisited regularly and adjusted based on community needs and systemic developments.

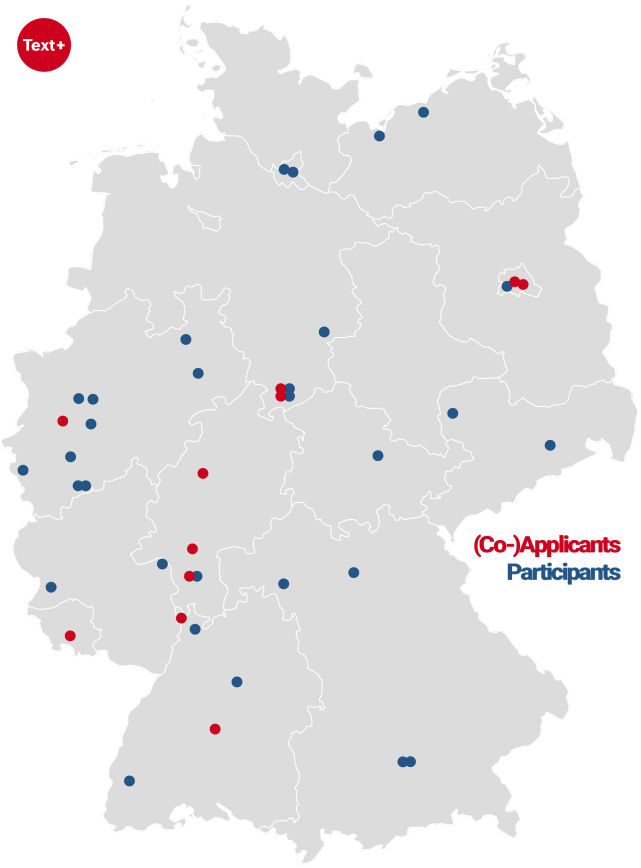
3. Consortium

Spokesperson Prof. Dr. Philipp Wieder, UGOE/SUB, is also a participant in NFDI4Biodiversity. Co-spokesperson Regine Stein, VZG, is also co-spokesperson of Base4NFDI and participant in NFDI4Culture. Participant Ralph Müller-Pfefferkorn, TUDD, is

also a participant in NFDI4Chem. Participant Kerstin Hoenig, DIE, is also a participant in BERD@NFDI in its potential second funding phase. Participant Johannes Mangei, HAB, is also a participant in NFDI4Memory. Participant Frank Krüger, HSW, is also co-spokesperson in NFDI4Datascience in its potential second funding phase. Participant Andreas Münzmay, UniPB, is also co-spokesperson in NFDI4Culture. Participant Christof Schöch, UniTR, is also a participant in NFDI4Culture. Participant Stefan Schmunk, h_da, is also co-spokesperson in NFDI4Memory.



3.1 Composition of the consortium and its embedding in the community of interest



As part of the preparations for the NFDI process, several Text+ partners have joined forces with a large number of professional associations in the Humanities as well as with resource providers to engage in a dialogue with scholars in the Humanities. These stakeholders participated in a series of three workshops titled Research-driven Research Infrastructures for the Humanities and Cultural Studies in Germany (Wissenschaftsgeleitete Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften in Deutschland), co-organised by the infrastructure initiatives CLARIN-D and DARIAH-DE, the Union of the German Academies of Sciences and Humanities, and the professional association Digital Humanities in German-speaking countries (Stein, 2018). As a result, four NFDI initiatives–NFDI4Culture, NFDI4Memory, NFDI4Objects, and Text+–were established, leading to the formation of four complementary consortia representing the Humanities. These consortia now jointly present themselves within the NFDI framework as the Humanities@NFDI group (NFDI e.V., n.d.b), based on a Memorandum of Understanding (MoU) prepared as early as 2019 and updated in 2020 (Brünger-Weilandt et al., 2020).

The initial composition along data domains represents types of data with a long tradition in the Humanities: language- and text-based Collections, Lexical Resources, and Editions. Reconfiguring Collections into “Library and Archive Holdings as Research Data” and “Language Corpora and Elicited Data” results in a refined composition into four data domains, supported by the task area Infrastructure/Operations, which also serves as the interface to the NFDI for all technical issues, particularly for basic service developments. All overarching tasks are coordinated in the task area Administration. In line with the main idea of the NFDI as a community-driven infrastructure, a joint coordination structure was implemented from the outset to reflect this co-operation: IDS with its expertise on text- and language-based science, has a focus on scientific coordination tasks, UGOE/SUB with its expertise in humanities infrastructure development has a focus on infrastructure coordination tasks. The two co-spokespersons in TA Administration are jointly responsible for the overarching integration within the consortium. In the first funding phase, IDS acted as the applicant institution and thus provided the spokesperson for the consortium. For the second funding phase, the applicant and spokesperson roles will be transferred to UGOE/SUB, focusing on Text+ infrastructure and operations management. The overall project coordination is a shared task between both institutions.

The emphasis on this dual management structure has proven successful for overall management and will therefore also be implemented in the task areas in the second funding phase. Each task area will be led by two co-spokespersons representing scientific and infrastructure management for the task area. To implement this, five universities previously in different roles will join as co-applicant institutions, with recognized professors serving as co-spokespersons: Humboldt-Universität zu Berlin with Prof. Vivien Petras PhD, Marburg University with Prof. Dr. Hanna Fischer, Saarland University with Prof. Dr. Elke Teich, Technical University of Darmstadt with Prof. Dr. Andrea Rapp, University of Tübingen with Prof. Dr. Angelika Zirker. In order to further strengthen the long-term perspective for the Text+ infrastructure and its sustainable operation, in addition the Verbundzentrale des Gemeinsamen Bibliotheksverbunds, the service center of the largest library network in Germany, will join with Regine Stein as co-applicant institution.

In phase I, tasks were organised around thematic clusters, which proved effective for launching the project and implementing the infrastructure. While working on the tasks, it became clear that a cross-cluster approach was needed to avoid pillarisation and enable interdisciplinary collaboration. Work on several topics is even more adequately addressed in cross-data domain activities, including Infrastructure/Operations resources, which is reflected in the establishment of permanent Text+ working groups, e.g., on the Text+ Registry and on Linked Open Data.

Hence, in the current proposal, thematic clusters play a secondary role, especially with the reconfiguration of Collections into two data domains. The emphasis now lies on collaboratively addressing, notably through working groups, overarching research data management challenges as well as the overarching topics as described in section 2.

Text+, as an open and community-driven consortium, offers several modes of participation to Humanities scholars at different stages of their career, to professional associations, and to institutions that want to join Text+ as data and service providers: Participant institutions make essential contributions to Text+ by providing their research data, services, and expertise. For example, GWDG, SAW, and UniBA are intensely involved in the provision of overall infrastructure components such as the Text+ website and its search and retrieval infrastructure. During the first funding phase, five new partners have joined the consortium as participants and will continue to contribute in phase II. Professional associations will go on to provide an organisational bridge to the communities of interest and support shaping the development of Text+ research data and services.

Text+ is a community-driven consortium that places the needs and perspectives of its scholarly communities at the center of its activities. Consulting and support services are a major cornerstone of community interaction. Text+ offers expert guidance on data management, legal and ethical issues, and technical implementation. These services are tailored to the specific needs of scholars working with language- and text-based research data and are delivered through one-on-one consultations, helpdesks, training events, workshops, and summer schools. Text+ also develops and disseminates Open Educational Resources (OER).

A specific focus of Text+ is the cooperation with the specialised disciplinary information services (Fachinformationsdienste; FID) (Text+ Consortium, n.d.f), which cover a considerable part of Text+'s communities of interest and a vast range of languages and cultures. A permanent Text+ working group FID Koop and the "FID/Text+ Jour Fixe" twice a year form the basis for intensified dissemination of and engagement with Text+'s portfolio in the respective communities.

For incorporating this continuously growing network of research institutions and individual scholars into Text+'s governance, a key mechanism is the structured participation of professional associations that nominate representatives to serve on the Coordination Committees (see 3.4). Elections and nominations are conducted transparently and inclusively, allowing for broad participation across career stages and disciplines, ensuring that the governance of Text+ reflects the diversity of its user base.

Text+ maintains a strong presence at academic conferences in the fields of linguistics, literary studies, digital humanities, and related disciplines (e.g. participating in 148 events, see appendix 5). This includes contributions to conference programs, the organisation of

workshops and panels, and the operation of information booths that provide direct contact points for researchers. In addition, Text+ hosts its own annual user conference and thematic workshops which serve as platforms for dialogue, feedback, and co-development of services and data offerings (see 138 events organised, see appendix 5).

To foster innovation and strengthen ties with the research community, Text+ has issued annual calls for collaborative projects (with 14 projects finished, 5 ongoing in 2025, and 14 in the review process for 2026), which invite proposals for new data sets, tools, and use cases that extend the consortium's portfolio. Additionally, early-career researchers are explicitly encouraged to apply to present at an early-career session at the annual plenary meeting. Travel support is provided to facilitate their participation and to promote visibility and networking opportunities within the community.

During phase I, Text+ developed and conducted a variety of training and educational formats, collaborating closely with the community. These activities can be broadly categorised into the following areas:

- Research Data Management: Integration of domain-specific expertise (e.g. Basisklassifikation), legal issues (e.g. Derived Text Formats), solutions (RDMO), standards and interoperability (e.g. standardisation series, Integrated Authority File GND), broad community engagement (FAIR February),
- Internal processes and services of the consortium (e.g. Helpdesk, Deployment Pipeline Portal, Service Discovery),
- Innovative topics that require adjustments in research data structure (e.g. Artificial Intelligence, Large Language Models),
- Infrastructure and service-related topics (e.g. IO-Lectures, Show&Tell series),
- Cross-consortial assessment of specific data types (e.g. data gathered from Social Media Platforms in the joint lecture and discussion series "Social Media Data – Show & Tell" with NFDI4Culture, KonsortSWD, and BERD@NFDI).

These efforts are based on a collaborative approach: topics initially treated internally are often opened to the wider community or conducted as joint initiatives (e.g., Humanities@NFDI, FID/Text+ Jour Fixe). Formats included 112 events (see appendix 5, field 1008, both on-site and virtual), and 42 published materials (see appendix 5, field 1005) comprising guidelines and blog articles.

This broad array of training and outreach activities is substantially developed further in phase II as detailed in the respective task area chapters (Measure 5 Communities). The integration of these activities into a comprehensive communication and training strategy will be coordinated in collaboration with the Communications Office.

Text+ actively supports the integration of training programmes into the curricula of its (co-)applicant and participating institutions. With the additional emphasis on university

partners – particularly with the new co-applicant institutions – a more direct involvement in curricular development has been initiated. Their engagement ensures that training in research data management (RDM), digital methods, and AI-supported approaches are structurally embedded in academic programmes at these institutions. While the formal adaptation of study programmes is a long-term process, due to accreditation requirements and a required cultural change, Text+ partners have already implemented a variety of curricular and co-curricular formats. These include the integration of the above mentioned topics into existing courses on the bachelor's and master's levels, workshops for doctoral candidates and postdocs, and continuing education for academic staff.

3.2 The consortium within the NFDI and the national academic research system

Text+ is a key contributor to the NFDI – with 40 interconsortial collaborations and 10 with the NFDI association (Eberl et al., 2025) – both as a disciplinary consortium and as an active participant in shaping the structures, services, and strategies of what is envisaged to become the OneNFDI. Text+ is also the key player regarding the development of a national infrastructure for the text- and language-related research community in Germany. Text+ accordingly considers one of its core tasks to collaborate closely with its peers in the academic research system in Germany. These collaborations are strategically integrated into Text+ processes and respond to the continuous development and changes in the broad landscape of text- and language-related research (see 3.1). Implementation-wise, the approach towards collaboration with the academic research system is also tightly integrated into the Text+ governance through the Coordination Committees (see 3.4), and it is, on various levels, incorporated into the research data management strategy of Text+ (see 4.1).

Text+ is actively contributing to the NFDI Association: It provided the spokesperson of the Konsortialversammlung (KV) for one election period (through the spokesperson of the first funding phase, Andreas Witt) and continues to participate in the KV's strategy group. Text+ was also involved in shaping the NFDI's shared vision and mission (NFDI e.V., n.d.d) and contributed extensively to the task forces (TF) of the NFDI Association (NFDI e.V., n.d.a), e.g. through the co-chairing of the TF Evaluation and Reporting (Thorsten Trippel, IDS) and input for the TF on the Mission and Vision, Strategy, and the TF Governance and Sustainability. Text+ is committed to further contributions to the NFDI association under the new dual leadership framework. This will involve, among others, linking the project's strategic themes to the evolution of the NFDI and coordinating Text+'s involvement, in particular through the co-applicant organisations and the co-spokespersons.

Through its partners, Text+ is actively involved in all five NFDI Sections (NFDI, n.d.c). Its contribution is distributed across different levels of participation and leadership, and is most

visible in several working groups. Selected examples include: Section Common Infrastructures: Co-chairing of the Working Group Overall Architecture (Philipp Wieder, UGOE/SUB); co-chairing of the Working Group Persistent Identifier Services (Sven Bingert, GWDG); co-chairing of the Working Group Long-term Archiving and Preservation (Thomas Stäcker, TUDa); participation in Working Group Identity and Access Management.
Section (Meta)data, Terminologies, and Provenance: Taskforce Metadata (Thorsten Trippel, IDS and EKUT). Text+'s work on metadata (see 4.2) is directly contributing to the NFDI wide discussion and aligned with it. Co-chairing of the Working Group Cookbooks, Guidance and Best Practice (Regine Stein, VZG). In addition, Text+ partners are engaged in Section Training & Education, participate in the working group RDM Helpdesk Network and actively helped setting up the Section Industry Engagement.
These engagements ensure that Text+ contributes to the development of core services, standards, best practices, and operational aspects across the whole NFDI.
Text+ will evolve its contribution in the second funding phase, thus aligning the contributions with its strategic themes (see 2.1), its overarching objectives (see 2.2), and the specific objectives of its task areas (see 5.0). A potential contribution worth emphasizing relates to the measures on Federated Research Data Infrastructure (for all data domain TAs, see 5.0, and infrastructure/Operations, see 5.5), which are closely related to the Section Common Infrastructure and the development of the NFDI as a distributed infrastructure based on a common architecture. However, Text+ will also consolidate its contributions, in particular to working groups, and will favour strategic contributions that help build a sustainable infrastructure as part of the OneNFDI. The respective decisions in such cases will be guided by the Text+ Steering Group (see 3.4). In particular with respect to the strategic theme "Language/text as dominant interface to data", new TFs or WGs are envisaged and could be instantiated with contributions and leadership of Text+.
The consortium maintains relationships with other NFDI consortia, contributing to a shared ecosystem of research data infrastructures. A comprehensive overview of all inter-consortial collaborations within the NFDI is published (Eberl et al., 2025). As one of four consortia organised as Humanities@NFDI Text+ has a long-standing tradition of collaboration that predates the formal establishment of the NFDI. This partnership is supported by regular meetings among spokespersons and coordinators, additionally by collaboration in the association for research infrastructures in the humanities and cultural studies (GKFI e.V.) (GKFI, n.d.). Together, the humanities consortia advocate for common needs and address shared challenges such as metadata standards, long-term preservation, and the implementation of FAIR data principles. In broader Social Sciences and Humanities (SSH) contexts, they often act as a unified voice.

Beyond the humanities, Text+ also collaborates with consortia from other domains. A notable example is the MoU with KonsortSWD from the Social Sciences, which focuses on the reuse of data sets and methods. In this context, Text+ initiated a working group on Social Media Data in collaboration with NFDI4Culture, BERD@NFDI, and KonsortSWD. This group launched a virtual workshop series titled “Show & Tell”, fostering interdisciplinary exchange and methodological innovation.

Knowledge transfer within and beyond NFDI is a core mission of Text+ and is embedded in its community-based governance structure (see 3.1 and 3.4). These activities foster knowledge exchange not only within Text+ but also across the NFDI and the wider academic landscape. Several Collaborative Research Centres (CRCs) at co-applicant and participating institutions serve as important hubs for knowledge transfer and infrastructure integration. These include: CRC 1252 – Prominence in Language (Cologne); CRC 1102 – Information Density and Linguistic Encoding (Saarbrücken); CRC 1288 – Practices of Comparing (Bielefeld); CRC 1391 Different Aesthetics (Tübingen); CRC 1412 – Register: Situational and Functional Aspects of Linguistic Knowledge (Cologne); CRC 1567 – Virtual Lifeworlds (Bochum); CRC 1646 – Linguistic Creativity in Communication (Bielefeld); CRC 1718 – Common Ground: Cognition – Grammar – Communication (Tübingen).

Many CRCs include INF projects that are responsible for developing and maintaining research data management (RDM) infrastructures. However, these INF projects are typically limited to the duration of the CRC’s funding period and are not sustained beyond it, which creates structural vulnerability. By collaborating with a sustainable infrastructure such as Text+, CRCs can ensure continuity for their RDM efforts. Text+ offers a stable framework that can absorb and extend the work of INF projects. This strategic alignment supports the long-term availability of research data and contributes to fulfilling the DFG’s requirements for good scientific practice, which call for data to be preserved and accessible for at least ten years. In addition to these CRCs the consortium has collaborated with 62 projects in the project application phase, particularly with regard to research data management, helping them to find the suitable Text+ service for their needs and align with best practices in data stewardship.

Numerous partners are also involved in CLARIN-D and DARIAH-DE, the German national nodes of the European Research Infrastructure Consortia CLARIN and DARIAH. These collaborations strengthen the consortium’s alignment with European standards and practices in the management and reuse of language- and culture-related research data. While these connections are rooted in national structures, they also serve as important gateways to broader international networks (see 3.3).

In addition, Text+ maintains relations with key national stakeholders in the research data management ecosystem. These include the German Institute for Standardization (DIN),

particularly the working group on language resources and language technology (DIN NA 105-00-05 AA) (DIN-Normenausschuss Terminologie (NAT), n.d.); Nestor e.V., the Competence Network for Digital Preservation; the network of Specialized Information Services (FID); RDM-DE (Research Data Management Germany); and the national Data Competence Centers (Datenkompetenzzentren). These collaborations support the alignment of Text+ infrastructure with national standards and contribute to the sustainability and interoperability of research data services in Germany.

Text+ considers universal access to common services as critical for the underpinning of data sharing activities within its communities and a complementary activity to the development of discipline-relevant tools and services. Text+ welcomes the Base4NFDI objective of harnessing the potential of existing networks to create common services to support RDM. It contributed to the application process for Base4NFDI from the outset: it was actively engaged in preparing a statement of the Consortia Assembly of the NFDI e.V. on basic services (Konsortialversammlung, 2022) and the organisation of a joint application by all consortia, including participating in working groups and organizing workshops to develop and refine the concept project. A Text+ co-spokesperson (Regine Stein, VZG) is co-spokesperson of Base4NFDI's Governance task area, and a Text+ co-applicant (DNB) hosts a Base4NFDI Service Steward. Text+ is committed to support the emerging Base4NFDI services. Overarching services provided by the NFDI will enable Text+ to use solutions that have a long-term and sustainable outlook, which requires stable hosting from trusted operators. Basic services assure interoperability with existing solutions both within other NFDI consortia as well as with European Open Science frameworks such as EOSC (see also 4.1). Text+ seeks to underpin Base4NFDI's participatory and community endeavour which is to be considered a founding principle of OneNFDI.

Text+ contributes to the Base4NFDI developments of IAM4NFDI and PID4NFDI, where GWDG is co-lead institution in each case. Furthermore, GWDG, JSC, and TUDD participate in Jupyter4NFDI and already provide sustainable resources through this project to Text+ and the NFDI in general. This engagement includes all previous and current phases of these three services. Based on its architecture, its service demand, and due to the increasing maturity of IAM4NFDI, PID4NFDI, and Jupyter4NFDI, Text+ has decided to re-allocate all its own development resources related to the respective services in the second funding phase and follow the "integration over implementation" approach. The respective resources are allocated in TA5 M3 "Platform and Core Services", activity 1 - integration of Base4NFDI services, and aim at integrating these services in the Text+ service portfolio.

Concrete results with respect to the initial integration include primarily work executed in the first funding phase in the task area "Infrastructure/Operations" through the activities "Basic

Identification Services" (within M2) and "Data Processing Pipelines" (within M5). The IAM solution for Text+, which is tightly integrated into the Text+ portal, is based on the AcademicID, which is one of the four IAM solutions that form the foundation of IAM4NFDI (IAM4NFDI Project Consortium, 2024). This ensures compatibility and long-term planning with respect to the integration of services and data sources. Handle-based Persistent Identifiers are widely used within the consortium, and Text+ benefits from the PID4NFDI offers regarding technical integration and support.

In addition to the aforementioned Base4NFDI services, which are closely integrated into Text+ architecture and service portfolio, Text+ partners are actively engaged with the following Base4NFDI services: Co-applicant VZG is through NFDI4Objects a partner in the integration phase of the TS4NFDI Terminology services, connecting VZG's mapping service Cocoda and the terminology register BARTOC. Participant TUDa is lead institution through NFDI4Ing in DMP4NFDI Data Management Plan services. Text+ has engaged with the service via its incubator offering and has entailed developing a dedicated DMP template.

Text+ continues to reserve resources to support structured ongoing communication with Base4NFDI. Text+ will participate in future Base4NFDI service incubator calls and pilots to prepare for the adoption of the services. Text+ will continue to closely monitor and connect to activities in Base4NFDI and will maintain its dedicated communication channel to Base4NFDI technical and community updates. Text+ will also reserve efforts in phase II to attend community workshops, calls, and training activities for each service.

To summarize, Text+ will during phase II – not least with regard to the expected transformation process towards a sustained operation of NFDI (RfII 2025; Wissenschaftsrat, 2025) – put special emphasis on

- continuing the concerted cooperation within the Humanities@NFDI group,
- further strengthening its already close collaboration with Base4NFDI and EOSC,
- and retain its well-established networks into its communities of interest on national and international level.

3.3 International networking

From the outset, Text+ has embedded its activities within a strong international framework, recognizing that sustainable research data infrastructures must be both nationally anchored and globally connected. This strategy is built on three core pillars: (i) collaboration with European Research Infrastructure Consortia (ERICs) and the European Open Science Cloud (EOSC), (ii) institutionalized international cooperation among partner institutions, (iii) and active participation in international standardization and RDM networks.

Text+ maintains close ties with the European ERICs CLARIN and DARIAH. These ERICs, along with the long-standing OPERAS initiative on the ESFRI roadmap, are directly

represented within Text+ through its partner institutions. UGOE/SUB represents DARIAH, IDS represents CLARIN, and MWS serves as the national node for OPERAS. This ensures seamless collaboration and mutual reinforcement of goals. To provide a sustainable structure for these activities, Text+ institutions founded the Association for Research Infrastructures in the Humanities and Cultural Studies (GKFI, n.d.). This association, chaired by Andrea Rapp (TUDa) and Stefan Buddenbohm (UGOE/SUB), serves as a legal and organizational framework supporting the ERICs. The coordination committee, comprising national coordinators for CLARIN, DARIAH, OPERAS, and external members, manages the service catalogue and coordinates in-kind contributions. Through this structure, Text+ facilitates national alignment of tools, services, and expertise, while strengthening its role within the broader Humanities@NFDI group.

These connections have allowed Text+ to integrate outputs from European working groups – on legal and ethical issues, metadata standards, and certification of data centers – into its own infrastructure. An example is the collaboration with DARIAH-IT, launched at the joint Text+ Community Plenary and European DARIAH Annual Event in 2025, hosted in Göttingen.

Text+ also contributes to the EOSC ecosystem. Project DataGEMS (DataGEMS Consortium, n.d.), funded as part of the EOSC portfolio, complements Text+'s national services. Additionally, partners are involved in the Social Sciences and Humanities Open Cloud (SSHOC) science cluster. Text+ has provided metadata on 117 services to the SSH Open Marketplace (Stefan Buddenbohm (UGOE/SUB) is a member of its Editorial Board), ensuring resources are accessible and reusable across European infrastructures. Technical interfaces developed by Text+ support metadata harvesting, interoperability, and integration with EOSC services. Text+ has contributed to the EOSC Federation Handbook (EOSC Federation, 2025a) and NFDI guidelines in collaboration with the Base4NFDI initiative, working closely with the NFDI Directorate and EOSC leadership. Text+ partner institutions, such as GWDG and JSC, are also involved in EU projects like AARC and the EOSC EU Node, showcasing technical expertise and an active role in shaping European infrastructure.

Beyond formal infrastructure collaborations, Text+ benefits from its partner institutions' international networks. These include participation in organisations like ALLEA, CENL, LIBER, IFLA, and EFNIL. A notable example is the initiative with Europeana and the RDA Interest Group “Collections as Data,” which combines datasheets for cultural heritage data with data management plans. Text+ experts also contribute to international standardization efforts through active participation in bodies such as the Research Data Alliance (RDA), the Text Encoding Initiative (TEI) (Text Encoding Initiative, 2025), and ISO. An example was the 2022 “National Libraries as Data” Dialogue Forum, co-organised by Text+, which brought

together ERICs and CENL to explore how national library data can be made more accessible for humanities research.

Institutionalized international cooperation is further supported by research projects involving Text+ partners. Text+ also supports internationalization through its infrastructure: by offering persistent identifiers, metadata standards, and long-term preservation services that meet international expectations, the consortium ensures its services are interoperable and sustainable beyond national borders.

Looking ahead, Text+ will continue strengthening its international engagement by deepening collaboration with ERICs, expanding contributions to EOSC, and supporting partners in international research initiatives. These efforts align with the strategic goals of the “Key Points for the Second Funding Period of the NFDI Consortia,” particularly task h) (NFDI Expert Committee, 2023), which calls for increased international visibility and integration of NFDI services. By embedding its infrastructure in international networks, Text+ contributes to the global advancement of open science and ensures that German research remains at the forefront of digital scholarship in the humanities.

3.4 Organisational structure and viability

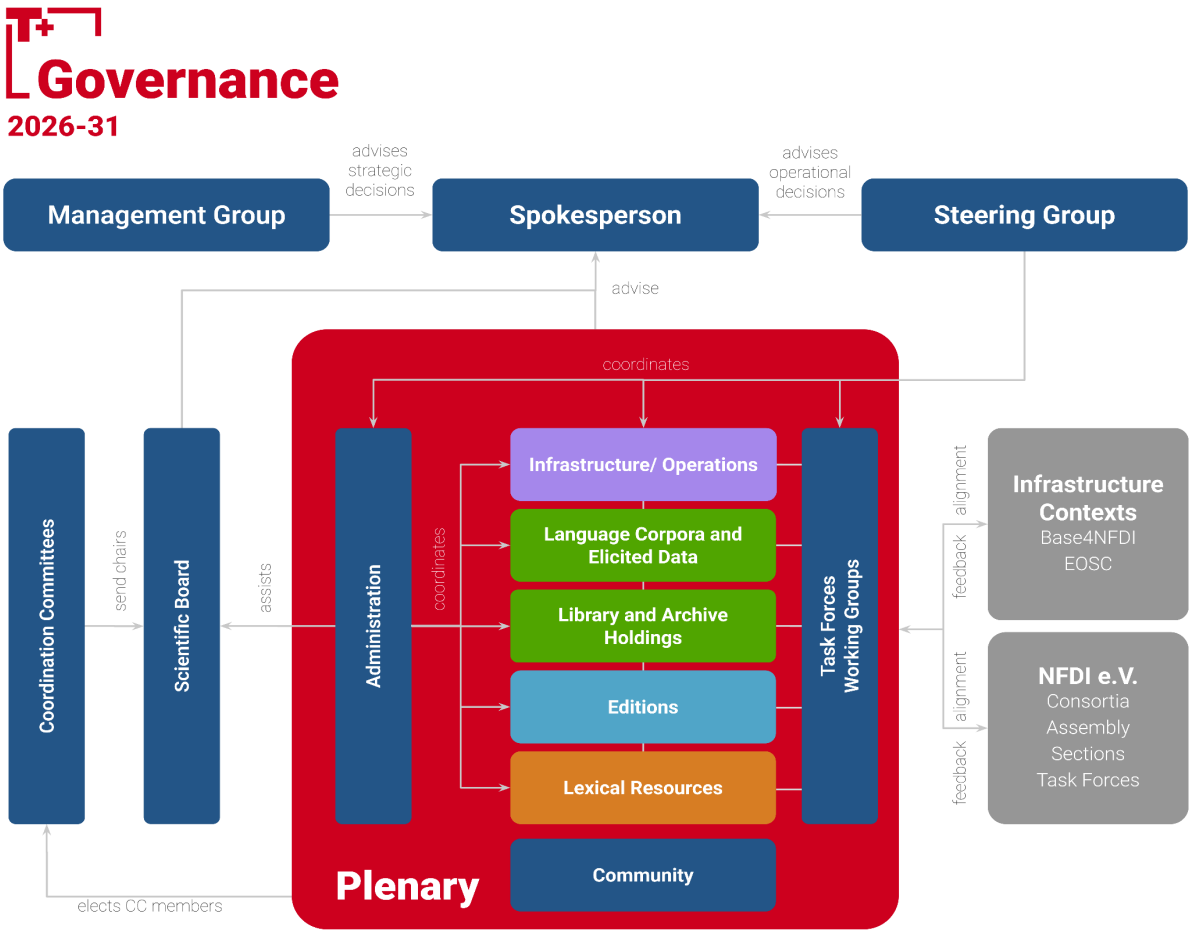


Figure 3.1: Organisational structure of Text+

The organizational structure of Text+ is from the outset designed to ensure effective governance, strategic oversight, and long-term sustainability in alignment with the goals of the NFDI in a developing and growing setting of its communities of interest. The governance model balances operational efficiency with broad community participation, while institutional commitments and sustainability measures ensure the viability of services and structures.

Text+ is governed by community-driven components, namely Plenary and Coordination Committees, on the one hand, and a two-tier system consisting of a Steering Group and a Management Group, on the other hand. The Scientific Board as interface ensures the implementation of community needs into the actual work. A concise overview of its main components is given here:

Plenary: An annual face-to-face meeting continues to be the central Text+ community event, allowing for orchestrated discussion, networking, and presentation of achievements to the consortium. It is announced publicly and open to all interested parties.

Coordination Committees: Each task area (except Administration) is accompanied by a Coordination Committee composed of members from the scientific community. These committees have the following functions:

- Advisory function: They review the project progress within their respective task area and provide recommendations for its further direction. This may include calling the consortium to extend their activities and developments in particular directions, but also providing funds from the flexible funds for these tasks which can be assigned to consortium members as they are willing to take up the task.
- Bridging function: They act as a link to the scientific community by engaging in professional societies, reporting there, and gathering feedback from the communities.
- Review function: They evaluate flexible fund projects within the framework of the project, prioritizing them to serve as a basis for the Scientific Board's decisions (see TA6 AD-M2 in the work plan).

Scientific Board: It revises and approves the recommendations of the Coordination Committees from an overarching perspective, in particular on the allocation of flexible funds. It also addresses overarching strategic issues.

Spokesperson: The spokesperson, appointed in accordance with the federal-state agreement BLV, heads the consortium, serves as the first point of contact, and represents the consortium externally.

Steering Group: Consisting of all Co-Spokespersons, this group serves as the decision-making body for steering and strategy. It oversees the progress of TAs and receives reports from working groups and temporary Task Forces. It also coordinates collaboration with other consortia and the NFDI association.

Working groups: These have proven to efficiently address cross-sectional topics across all Task Areas and are responsible for creating cross-TA deliverables and products. They are supported by the office, which ensures overall consistency and communication.

Management Group: The Management Group, comprising leading representatives of co-applicant institutions, is responsible for overarching strategic alignment and sustainability. It advocates for the consortium and engages with high-level stakeholders. Members are formally delegated and not subordinate to the Steering Group, ensuring independence. It exercises an oversight function and approves key planning documents.

Consortium as part of the NFDI Association: According to the NFDI Association statutes, Text+ is both a DFG-funded “Konsortium gemäß BLV” and, as part of the Association, a “Konsortium gemäß Satzung”. The transition path to a sustained structure in the Association will evolve within the second funding phase together with all consortia.

The governance structure has remained stable overall, with adjustments made primarily in response to personnel changes such as retirements or institutional transitions. The core structure and principles of governance as outlined in the original proposal will be preserved. They were successfully implemented during the first funding phase and are running smoothly. The additional data domain can seamlessly be incorporated into the governance, proving the model’s flexibility.

The financial planning for the upcoming funding phase builds upon the structures established in the previous proposal. Overall, the spending of funds has remained closely aligned with the original financial plans, with only minor deviations, primarily related to adjustments in the distribution among consortium partners to ensure a fair and effective allocation of resources based on project needs. However, in light of inflation and other cost-increasing factors, the real value of the available budget has decreased compared to the previous phase. This, combined with insights gained during the past funding period, has prompted a strategic reassessment and restructuring of how funds will be distributed in the future. A key change in phase II is the reduction in the number of partners receiving continuous, fixed funding. This allows for greater flexibility and responsiveness to recent and emerging developments within the consortium and the broader research data landscape. A core group of partners, including the co-applicants, will continue to receive a basic budget. This funding will ensure the maintenance of essential services and the execution of foundational tasks that are critical to fulfilling the work plan. The instrument of flexible funds, which has been designed and was implemented successfully in phase I, will be adapted in phase II based on the strategic themes and the project’s objectives (see 2.1.4).

The viability of the organizational structures throughout the second funding phase of the NFDI is ensured by the strong commitment of the (co-)applicants and all participants of

Text+. This commitment is evident through substantial allocation of institutional funds and provisioning of research data and services, as outlined in the Letters of Commitment. Partners will be systematically evaluating and integrating the needs of their communities into their strategic planning (see 3.5).

To ensure the demand-oriented further development of sustainable services and structures, the consortium has implemented several governance measures:

- Regular feedback loops with the scientific community and stakeholders.
- Continuous monitoring and evaluation of service usage and effectiveness.
- Strategic planning sessions to align services with emerging needs and trends.

Training and outreach activities, including workshops, documentation, and community engagement, support the broader goal of increasing service adoption and promoting best practices in research data management.

Text+ applicant institutions will advocate for a general NFDI business model based on the open access principle. The consortium believes that scientific data resulting from publicly funded research should be freely available to the public and researchers. This approach is essential for creating new research opportunities and contributing to the preservation of cultural heritage.

Text+ greatly appreciates the latest impulses from the Wissenschaftsrat (Wissenschaftsrat, 2025) and the Rat für Informationsinfrastrukturen (Rat für Informationsinfrastrukturen, 2025) towards permanently incorporating the NFDI into the German research system. While the second funding phase will in the first place focus on consolidating the consortium's work and governance structures, Text+ will actively participate in the process to ensure the NFDI's long-term viability and help shaping the OneNFDI, just as the consortium has contributed to cross-consortium collaboration within the association during the first funding phase.

3.5 Operating model

The operating model of Text+ defines the structure, processes, and governance needed to provide distributed services for text- and language-based research. By the end phase I, Text+ will have an operational service portfolio already in strong use by its community of interest. Phase II, therefore, focuses on stabilising operations, strategically evolving the service portfolio while safeguarding relevant existing services, and supporting the long-term integration of Text+ into the NFDI, as well as its connectivity with other European research data infrastructures and, in particular, EOSC.

The operating model is structured to meet the needs of both users and providers. Users are offered reliable access to data, the means to upload and archive data, and a portfolio of domain-specific tools and services, all accessible through the Text+ Portal (see 4.4). The

service portfolio is continuously adapted, based on user requirements collected through established consultation processes within the consortium. On the provider side, Text+ Centers constitute the backbone of the operating model regarding data access and preservation. These centers combine data storage and provision, technical services, and domain-specific expertise across the various data domains supported by Text+. Their institutional mandate to support public research ensures that the data and services they provide remain free of charge and are sustainably available to the research community. Core to the service portfolio of the Text+ Centers and the Text+ operation model is the data depositing service (Text+ Consortium, 2025d). Supported by the Text+ Helpdesk, data owners from the community of interest are guided through all steps towards the FAIR integration of their data into the Text+ Data Space (Barth et al., 2023a). For service providers, Text+ maintains a structured service management approach with defined procedures for service onboarding, operation, and offboarding. A concise governance structure (see 3.4) ensures that decisions regarding the portfolio are made based on the experience of the Co-Spokespersons considering scientific, strategic, and economic impact. The service portfolio includes technical, infrastructural, and non-technical components, such as consulting, helpdesk support, and training resources. Services are catalogued in the SSH Open Marketplace and data in the Text+ Registry using machine-readable, structured metadata to ensure interoperability within NFDI and beyond. The portfolio clearly distinguishes between services developed with Text+ funding, services provided by partners outside the project context, and external services recommended by Text+ but not included in its operating model.

All participating institutions provide substantial in-kind resources, including infrastructure operation, technical development, and data stewardship. Services will continue to be provided free of charge in phase II. There are no plans to introduce user fees in the near future. The long-term financial sustainability of the NFDI depends on its evolving operational framework and possible shared national cost models, in which Text+ is actively involved (see 6.2 for further comments).

Service management follows established guidelines adapted from CLARIN, DARIAH, and NFDI-wide practices, as well as IT Service Management standards introduced by Text+'s IT service providers JSC, GWDG, TUDD, and VZG. Defined roles include Portfolio Manager, Service Owner, and Service Provider. As the services mature, measures such as service level agreements (SLAs), monitoring, and analytics will be implemented to ensure operational stability and quality assurance.

The technical infrastructure integrates services such as the Text+ Registry, Federated Content Search and domain-specific tools. Cloud-based project-internal services are already in place and will be maintained, e.g. for event management (Indico), document

storage and exchange (NextCloud), and DevOps (GitLab). The second phase will prioritize the further integration of Base4NFDI services, including PID4NFDI, IAM4NFDI, Jupyter4NFDI, and nfdi.software (see 3.2). This will build upon joint development within the NFDI, with the aim of partially sharing the operation of a subset of NFDI services. The infrastructure will support both Text+ domains and broader NFDI use cases by leveraging partnerships with NFDI consortia such as NFDI4Culture, NFDI4Memory, and NFDI4Objects (see 3.2).

Text+ governance ensures coordinated development and operational responsibility through the Scientific Board, domain-specific Coordination Committees, the Management Group, and specific task forces. Partners contribute infrastructure, personnel, and resources to warrant alignment with community needs.

Text+ actively contributes to the NFDI and the development of OneNFDI, including shared infrastructure, metadata standards, training, legal frameworks, and operating models. Furthermore, Text+ maintains established connections with European infrastructures, such as CLARIN ERIC, DARIAH ERIC, OPERAS, SSHOC, and EOSC. This ensures alignment with European standards and access to interoperable services across borders.

4. Research data management strategy

During the first funding phase, Text+'s research data management (RDM) strategy focused on the three grand challenges (see 2.1.2), establishing “the fundamental path towards the integration of data, applications, tools, and services into an interoperable, standards-based infrastructure”. Text+ continues to implement this strategy, offering a stable, FAIR (see 4.3), and reliable infrastructure that provides access to rich data sets and a growing service portfolio for its user community. The RDM strategy was refined during the first funding phase, as necessary, for example, through the introduction of the Text+ Data Space concept (Text+ Consortium, 2025c) and the ongoing development of Text+’s LLM service (Text+ Consortium, 2025e).

The results of the first funding phase provide a stable foundation for the evolution of the RDM strategy guided by the strategic themes (see 2.1.2) for the second funding phase. This will enable Text+ to (i) integrate more data centers and data sources, (ii) enhance the overall data quality, (iii) develop guidelines regarding sovereignty, legal aspects, and data ethics, (iv) support further scientific disciplines, (v) implement community-driven use cases, (vi) refine the user experience for future, sustainable operation, (vii) offer more well-integrated services, and (viii) strengthen RDM-specific contributions and collaborations within the NFDI and towards EOSC.

4.1 Scientific relevance and quality of the measures

Text+ builds on the established mechanisms of the initial funding phase and remains fully committed to the FAIR and CARE principles, which are particularly relevant in the context of emerging challenges posed by LLMs. These challenges include the opacity of model architectures, the loss of epistemic traceability, and the increasing dependency on commercial infrastructures that reconfigure the norms of access, authorship, licenses, and scholarly validation. The new project phase introduces two significant areas of development: First, the project aims to further align infrastructure components with the evolving needs of the research communities, promote cross-domain collaboration, and enhance interoperability and reuse across the Text+ data domains. Second, the growing importance of LLMs introduces a new thematic focus area that will be addressed both critically and constructively. Text+ will contribute to the responsible integration of LLMs into research workflows by providing curated training and evaluation data, promoting transparency in model development, and ensuring that these technologies adhere to ethical and legal standards. These efforts will further professionalize RDM within the addressed communities and reinforce the relevance of Text+ as a trusted and future-oriented infrastructure partner. In assessing the success of its implementation, Text+ applies a demand-driven, community-centered approach that continuously monitors and adapts its data and service portfolio.

Both the relevance and quality of the RDM strategy are evaluated through community feedback, the adoption of FAIR-aligned workflows, and the extent of integration into national and international infrastructures. A key pillar of the Text+ infrastructure strategy is the network of certified data centers which play a central role in ensuring the quality, sustainability, and accessibility of research data. These centers are strategically distributed and technically mature, allowing them to support a wide variety of data types, formats, and disciplinary needs. Their strategy is twofold: on the one hand, they guarantee long-term preservation and accessibility through persistent identifiers, certified repository standards (such as CoreTrustSeal or nestor), and adherence to community-specific metadata schemas; on the other hand, they actively contribute to the evolution of Text+ by integrating new data sources. Through continuous engagement with the research communities and active involvement in national and international standardization initiatives, the data centers ensure that the infrastructure remains robust, scalable, and aligned with current best practices in RDM.

4.1.1 Results of the successful implementation of measures in phase I

A successful outcome is the Text+ Data Space, which establishes and enhances the findability, searchability, and reusability of humanities research data. A data space, by

definition (Glombiewski et al., 2024), integrates data from different sources by relying on existing mapping and matching techniques, while maintaining data sovereignty at the source level. The Text+ Data Space comprises the Text+ Centers – data and competence centers providing RDM services – the Text+ Registry, the FCS, as well as the Text+ service portfolio provided by the participating Text+ Centers, the IT service providers within Text+, and Base4NFDI. The Text+ Centers form a distributed environment for managing existing data and new data. By distributed data management, all aspects of complex legal states of the data, diversity of data types, and disciplines can be addressed. The Text+ Data Space has been successfully implemented across all data domains (Genêt et al., 2023a). All services of Text+ are accessible via the Text+ portal (Text+ Consortium, n.d.b).

The Registry and the FCS are two central services that provide access to research data. The Text+ Registry (Text+ Consortium, n.d.c) is a versatile and flexible system for describing and cataloguing diverse entities and resources. The Federated Content Search (Text+ Consortium, n.d.e) is the central application in Text+ for searching directly in distributed research content. It supports searching in (often extensively annotated) data sets of Text+ Centers. The FCS is being developed in close cooperation with European partners, enabling access to resources already available there and vice versa. The modular concept, consisting of Text+ Centers, is scalable and open to new partners as well as new data domains. Working groups have also extended the portfolio of Text+ by providing data management services (Text+ Consortium, 2024b) to third parties (see also 4.3).

4.1.2 Evolution of the research data management strategy in the second funding phase

Over the first 15 months of the second project phase, Text+ will actively solicit new user stories that are closely aligned with the specific services and offerings outlined in the objectives of phase II. In contrast to the broader, more exploratory user stories collected during the initial funding phase, this new round of user engagement will focus on targeted use cases that directly address the applicability, usability, and potential of the Text+ infrastructure. These user stories are expected to provide concrete feedback on existing services and offer specific insights for their further development. In doing so, they will not only help validate the relevance of the infrastructure in diverse research contexts but also contribute strategically to its refinement. Moreover, the selected user stories will serve as multipliers by showcasing best practices, encouraging wider adoption across disciplinary communities, and fostering deeper integration of Text+ into the everyday workflows of researchers.

The evolution of Text+'s RDM strategy is guided by the strategic themes (2.0), ensuring that the planned measures contribute to the professionalisation of RDM. These measures are

grounded in clear criteria for assessing their relevance and quality and are designed to address the evolving needs of the scientific communities. The following strategic priorities, together with the relevant strategic themes (ST, see section 2.1.3), define the areas of implementation for the RDM strategy in the second funding phase.

The expansion of the Text+ Data Space will be pursued through the integration of additional centers and new data sources (ST2, ST4). Particular emphasis will be placed on strengthening and enhancing collaboration with research consortia, such as CRCs and ERC-funded projects (see 3.2 and 3.3), to position Text+ as a hub for managing, preserving, and making research data accessible within the relevant communities of interest. The measure “Federated Research Data Infrastructure” contributes to this category in each TA.

In parallel, enhancing the quality and usability of the Text+ Data Space remains a core objective (ST1, ST2, ST4). The preparation of data in various formats, including those suitable for LLMs and derived textual resources, will facilitate greater accessibility, reusability, and interoperability of data. The measures “Enabling Text and Language Data for AI” (TA1-4) and “Implicit and Explicit Structures” (TA1-4) contribute in particular to improving data quality. Additionally, TA5 features a specific focus on user experience and contributes directly to ST4.

Regarding legal frameworks, data ethics, and sovereignty, Text+ will continue to provide legal advice and support on copyright, licensing, and ethical issues through the Text+ Helpdesk (Text+ Consortium, n.d.g) (ST3, ST4). The legal and ethical challenges related to the development and use of LLMs within Text+ are expected to become a central focus within this category (see also (Text+ Consortium, 2025a) for an initial analysis).

Text+ will also expand its collaboration with scientific disciplines to engage with relevant research communities and further develop its service offering in alignment with OneNFDI (ST1, ST3). This includes, among others, the field of legal studies, as also recommended by the evaluation of the interim report, and specifically criminology, where Text+ has already received a letter of support from the Institute of Criminology at the University of Tübingen, expressing interest in using Text+ services and gaining access to specialized language models, for example, trained on legal texts and dissertations.

The starting point for capturing and managing the needs of the community of interest within Text+ are the user stories (Text+ Consortium, n.d.d) (ST1-4). During phase I they formed an foundation (Feidicker et al., 2024) and evolved into use cases, elaborating on the “description of needs” for services or processes within the Text+ Data Space (Genêt et al., 2025).

The evolution of the RDM strategy in the categories related to the consolidation of services and the development of RDM contributions (ST2, ST3, ST4) are described in section 4.4.

The collaboration within the NFDI will, as indicated above, be heavily influenced by the services offered in each of the TAs (see 5.0) (ST1, ST3, ST4). This also holds true for potential contributions of Text+ to the EOSC Federation (e.g., through an NFDI ESOC National Node (EOSC Association, 2025b)).

The relevance and quality of the Text+ RDM strategy are assessed based on a set of clearly defined success indicators that ensure both scientific alignment and operational effectiveness. These criteria evaluate progress in implementing the strategy, its responsiveness to community needs, and its contribution to national and international research data infrastructures. Concrete success indicators are defined both for measuring progress toward the overarching objectives of the second funding phase (see 2.2) and for the individual task areas (see 5.0).

4.2 Metadata standards

Metadata standardisation is a foundational element across all Text+ data domains. The consortium builds on established metadata standards and infrastructures while integrating community-specific requirements and fostering interoperability at both national and international levels.

Applied Standards: Text+ partners apply a wide range of national and international metadata standards depending on data type, research tradition, and technical infrastructure. These standards include: Dublin Core (DCMI Usage Board, 2020), DataCite (DataCite Metadata Working Group, 2024), MARC21, ISO 24622-1 and ISO 24622-2 (CMDI), ISO 24619 (Persistent Identification of Language Resources), DCAT (Albertoni et al., 2024), and LIDO (ICOM-CIDOC LIDO Working Group, 2021). As bibliographic metadata intended for library catalogues is usually not sufficient for research data, richer metadata schemas, e.g. defined according to CMDI (Component Metadata Infrastructure, standardised in ISO 24622-1 and ISO 24622-2), are extensively used for language data to flexibly describe complex language resources. As TEI is widely adopted and contains metadata in its header, TEI-based metadata remains crucial.

Consortium-Wide Strategy and Metadata Model: Text+ has implemented a harmonized, layered metadata strategy, which also reflects the current state of metadata standardization discussions within the NFDI (Henzen et al., 2025). This includes:

- Core layer: A shared set of metadata fields (in ISO 12620 terminology *data categories*) is required across all data domains, aligned largely with DataCite, to support FAIR principles and semantic interoperability in a domain-independent and cross-disciplinary manner.
- Domain-specific layer: Extends the core to accommodate specific needs of data according to an underlying data typology (e.g., editions, lexical resources, collections).

- Center-specific layer: Allows repositories to incorporate additional data categories as required by their local systems or data characteristics.

This model ensures that metadata remains flexible and expressive while enabling integrated search and discovery in the Text+ Registry, which supports browsing across harmonized metadata and utilizes authority files, such as the Integrated Authority File (GND), to facilitate entity linking.

Contributions to Metadata Standards and Common NFDI Architecture: Text+ contributes actively to metadata-related standardization processes in a layered, bottom-up manner:

- TA-based standardization by defining a TA-specific metadata model for each data domain to provide data set descriptions, based on the requirements of the Text+ centers for their specific data (Marčić et al., 2023).
- Consortium-internal standardization by defining a core metadata set (see Text+ common metadata in the ingest process; Genêt et al., 2023a; Text+ Task Area Editions, 2024) based on the data domain metadata elements and including controlled vocabularies as value schemas for individual metadata categories, also achieving cross-domain harmonization.
- NFDI-wide metadata standardization within the NFDI Section Metadata, Terminologies, Provenance, with contributions to their task force on developing a core metadata standard.
- National standardization within the German Institute for Standardization (DIN) in the working group DIN 105-00-06 AA (“Language Resources and Language Technology”).
- International standardization within the International Organisation for Standardization in ISO TC 37 SC 4 (Language Resource Management).

The NFDI Section Metadata, Terminologies, Provenance promotes a cross-disciplinary core metadata model, echoing the layered metadata model used in Text+, and supports interoperability across consortia. Joint workshops of the NFDI section, in which Text+ partners were part of the organising group, have helped define expected outputs of the section, including:

- The recommendation to register repositories with re3data.org (re3data.org, 2013), with the requirement for regular updates and maintenance of repository entries.
- The definition of a cross-consortial metadata core drawing on DataCite, DCAT, and schema.org (Schema.org, n.d.).

Text+ anticipates that its current metadata models – already strongly aligned with DataCite – will readily comply with future NFDI recommendations. Internal mapping projects (e.g., CMDI to DataCite/schema.org) have laid the groundwork for smooth adoption.

Interoperability, Reusability, and Registry Integration: To integrate diverse repositories and foster reuse, Text+ has developed interoperability models that map existing metadata schemas to the internal Text+ model (see e.g. Genêt et al., 2023b). Metadata is harvested via OAI-PMH (Open Archives Initiative, n.d.), RESTful APIs, and other interfaces and is, where needed, enriched by references to authority files, also utilizing the GND agency of Text+ (Text+ Consortium, n.d.a). The facets and information displayed for data sets in the Text+ Registry are also based on the Text+ core metadata schema, with continuous consolidation efforts improving metadata consistency. This ensures both discoverability and depth of documentation by applying the specific data categories of the Text+ centers.

Further Development with regard to Metadata Standards: Text+ will continue to shape the common metadata infrastructure of the NFDI by implementing future cross-NFDI metadata recommendations, enhancing semantic interoperability through knowledge graph alignment, contributing training and capacity building, especially around GND-based entity linking, and coordinating with other NFDI consortia to feed into metadata sustainability, reusability, and machine-readability. In doing so, Text+ aims not only to maintain but expand its leadership role in metadata standardization, providing both a rich testbed and a stable foundation for the NFDI's overarching infrastructure goals.

4.3 Implementation of the FAIR principles and data quality assurance

Within the consortium, FAIRness is not only defined by adherence to technical standards but also by a commitment to the specific needs of the linguistic, literary, and cultural research communities. These needs include the handling of diverse data types and annotation formats, as well as navigating legal constraints, particularly in relation to copyright and the protection of personal data in spoken language corpora.

The Text+ Data Space provides the conceptual and technical framework for FAIR-aligned data management and is complemented by consulting services, a helpdesk, and community engagement activities that promote best practices and ensure responsiveness to evolving scholarly requirements.

Findability is ensured through the Text+ Registry which aggregates rich metadata from all participating centers. The metadata includes persistent identifiers such as URNs, DOIs, and handles, following DIN ISO 24619:2018. Accessibility is addressed through a tiered access model that reflects the varying degrees of openness and legal restrictions associated with different data sets. This model provides open access to unrestricted data, controlled access via identity and access management services such as IAM4NFDI, a Base4NFDI service, and federated content search mechanisms that enable legally compliant access to sensitive or protected data through derived text formats. Interoperability is achieved through the use of standardized metadata schemas and the integration of authority files such as the GND.

The Federated Content Search (Körner & Eckart, forthcoming) is being extended to support additional resource types, including lexical data via LexFCS (Eckart et al., 2023; Körner et al., 2023). Moreover, the Registry and FCS are designed to interface with external infrastructures such as the SSH Open Marketplace and nfdi.software, thereby enhancing cross-consortial interoperability. Reusability is supported through clear licensing, comprehensive metadata, and adherence to established community standards. Text+ also offers depositing services to researchers outside the consortium, guiding data quality and documentation to warrant alignment with FAIR principles.

Beyond the FAIR framework, Text+ places strong emphasis on additional quality criteria. These include repository certification through CoreTrustSeal (CoreTrustSeal, n.d.) and nestor Seal (according to the DIN 31644 standard), (German National Library (DNB), n.d.b), legal compliance with copyright and data protection regulations, semantic enrichment through the use of authority files and controlled vocabularies, and comprehensive user support through documentation, training, and outreach. Data quality is further ensured through initial review and curation by data centers, metadata validation and enrichment, and ingest procedures aligned with the OAIS reference model (ISO 14721:2025).

Looking ahead, Text+ plans to develop automated tools to match researchers with appropriate data centers based on their specialization, expand quality metrics and validation tools grounded in existing standards, and enhance user feedback mechanisms to support continuous improvement. The consortium maintains a decentralized model of data stewardship, with each data center responsible for its repository and long-term archiving strategy. Access policies are defined at the center level following legal agreements and institutional frameworks, and IAM integration will be expanded to streamline access to restricted resources.

Text+ is fully compliant with the General Data Protection Regulation (GDPR) and national data protection laws. Personal data, particularly in spoken language resources, is handled with strict access controls, anonymization where appropriate, and legal review prior to publication. IAM services provide an additional layer of protection by ensuring that only authorised users can access sensitive data.

Finally, cooperation projects play a key role in integrating external resources and refining ingest workflows (see 4.1.1). These projects contribute to the expansion and diversification of the Text+ portfolio and help ensure that external contributions meet the consortium's FAIR and quality standards.

4.4 Services provided by the consortium

Over the years, the various stakeholders involved in Text+ (see 3.1) have developed a wide range of services for their communities of interest through third-party funded projects or as

part of their institutional missions. Based on the Text+ Service Policy (Buddenbohm & Weimer, 2024), selected services have been onboarded into the Text+ service catalogue, which is an integral part of the Text+ portal. Currently, the catalogue lists 79 (depending on the specific counting method) services (Text+ Consortium, 2024b), representing Text+'s current service offering, including repositories, FCS-endpoints, APIs, and services for search, annotation, analysis, and publishing. The Service Policy clearly states the criteria for integrating services into the catalogue. These include (i) a service design based on the Text+ architecture (see figure in 5.5), (ii) integration into the Text+ portal as the central access point for users to find, access, and deposit data and services; and (iii) adherence to governance processes to ensure balanced decision-making with regard to technical issues and resource usage. The Text+ Service Policy also describes the service management approach that has been implemented to manage the service lifecycle. It includes processes for onboarding, service operation, and offboarding/termination.

To enable sustainable operation, Text+ is implementing a strategy comprising multiple building blocks. The first of these is the Text+ centers (see the operating model in 3.5 for more details). Each individual Text+ center generally provides and sustains resource-specific services, such as query endpoints for specific corpora and dictionaries, online presentations of editions, dictionaries, and corpora, according to their institutional mission. Through the centers, a certain service level is guaranteed, independent of any NFDI funding. In particular, long-term availability and accessibility of research data is ensured. The second building block is the Text+ Architecture. This has been developed (Text+ Consortium, 2025b) and is updated in line with the results of the EOSC and the NFDI WG Overall Architecture. Furthermore, participant GWDG is actively contributing to these endeavours to ensure that Text+ requirements are met and that Text+ contributions are valuable. In phase II, TA5 will define a concise "NFDI-compliant" core platform through the "Platform and core services" measure (see 5.5). Wherever possible, renowned IT service centers (JSC, GWDG, TUDD and VZG) will integrate the respective core platform services into their scalable, secure infrastructures. The platform will also facilitate the transfer of services between operators. One specific example is the planned transfer of the Text+ Registry to participant VZG during phase II. The third building block involves the strategic integration of, and contribution to, sustainable service developments, particularly those offered through Base4NFDI projects (see 3.2). Wherever feasible, this approach is favoured over the implementation of new services within Text+. Text+ activities realized through services, such as IAM4NFDI or Jupyter4NFDI, have therefore been transformed into smaller, less resource-intensive integration tasks in phase II.

In phase II, the Text+ service catalogue will evolve with the integration of further services, the continued development of services, and the development of new services by the data

domains and Infrastructure/Operations, according to the work programme and through flexible funds (see AD-M2). This final category will include automated annotation tools (TA1), NLP pipelines and LLM-based enrichment services (TA2), a lexico-statistical resource generator (TA3), a repository of reusable annotated corpora, pre-trained models, and AI tools (TA4), and LLM infrastructure services tailored for Text+ (TA5). These services, as well as upgrades to existing services, will be developed using NFDI funding in phase II. While the aforementioned building blocks, in combination with the operating model, guarantee a solid Text+ service offering, it is clear that the number of available services, the depth of integration, and long-term maintenance, as is generally the case for the NFDI, rely on continuous investment.

The services are monitored continuously on two different levels: technically, with respect to their availability and in terms of their usage. Results relating to the latter are reported as indicators field 4001–4008 and field 5001–5109 in the data sheet (see appendix 5). This data also informs the coordination of activities relating to promotion, dissemination, and training. These activities are carried out through the common measure M5 Community for the data domains and through the Text+ Communication Office (AD-M5) in general. Furthermore, in phase II, Text+ will extend this set of measures and provide training on service usage and info days regarding the Text+ service catalogue to the projects it supports (currently 66).

The Text+ Architecture and the Text+ Data Space are, in general, designed and implemented according to current development best practices, and the services are well-integrated in scalable infrastructures. This makes the question of service scalability mainly a matter of available infrastructure and licence costs, issues that can be managed at the governance level. However, the Text+ Service Policy will be extended and updated in phase II to introduce Service and Operational Level Agreements and to put greater focus on the Technology Readiness Levels of the various services in the catalogue.

4.5 Impact of changes of external conditions / constraints

Since the inception of Text+, a number of external developments have had significant impact on the landscape of research infrastructures in general, the services that Text+ develops and provides, and the evolving needs of its scientific communities. These include technological shifts, structural changes within the broader NFDI ecosystem, and the evolution of the EOSC, as well as legal and regulatory developments. This section highlights the most relevant developments and outlines how Text+ responded to them during phase I as well as how it intends to respond in future.

Technological Shifts: LLMs: The emergence and widespread availability of LLMs, particularly since ChatGPT was released in late 2022, has presented the Text+ consortium

with both opportunities and challenges. Although the community was already familiar with the use of language models, the scale, accessibility, and generative capabilities of current LLMs represent an unexpected and disruptive shift. In response, Text+ has established a cross-task-area working group to explore the integration of LLMs and Retrieval-Augmented Generation (RAG) technologies into its services. The group is investigating how LLMs can support specific research tasks, how Text+ data can be used to fine-tune or enrich such models, and how to address legal and ethical challenges, particularly those related to copyright and licensing, which currently limit the use of original data for training or publishing models. One strategy under consideration is the use of derived text formats which may be better suited to machine learning applications. A milestone is the Text+ LLM Service, which is integrated into the Text+ Portal and offers i.a. access to a variety of open and commercial large language models (see also 4.4) (Text+ Consortium, 2025e).

Text+ has already taken on the task of informing its community about LLMs and is actively creating opportunities for discussion and knowledge exchange. The 3rd Text+ Plenary, which took place in October 2024, was therefore dedicated entirely to this topic. Moreover, Text+ is engaging with national and international initiatives in the context of LLMs, including the DHd working group Applied Generative AI in the Digital Humanities and the European Language Data Space (LDS) which aims to foster the development of European language technologies.

Structural Changes in the NFDI: Base4NFDI: A significant structural development within the NFDI was the launch of the Base4NFDI initiative in March 2023. From the outset, Text+ as a project as well as individual co-applicants or participants have been involved in Base4NFDI, contributing to the conceptual and technical development of basic services. Selected areas of involvement include IAM4NFDI (authentication and access management), PID4NFDI (persistent identifiers), Jupyter4NFDI (interactive computing environments), and nfdi.software (software sustainability and metadata). Text+ also plays a key role in Base4NFDI governance, with one of the co-spokespersons coming from the consortium. Furthermore, service stewards for key infrastructure components, such as federated storage and semantic interoperability, are co-located with Text+ partners. This aligns Base4NFDI services closely with the needs of the Text+ community.

Evolution of the European Open Science Cloud: The launch of the EOSC EU Node in October 2024 marked a significant milestone for the European Commission, indicating a major change in the structure, governance, and operational model of this pan-European research infrastructure. The EOSC EU Node is the core node of the evolving EOSC Federation (EOSC Association, 2025a), which aims to integrate infrastructure providers and national research data initiatives, also known as thematic nodes, that provide data and services for specific scientific domains. In March 2025, the first 13 candidate nodes were

selected to form the extended core of the EOSC Federation. The NFDI EOSC National Node (EOSC Association, 2025b), among these, represents the NFDI's contribution in terms of use cases and services, potentially paving the way for larger-scale integration of NFDI services into the EOSC Federation in the future. Text+, which co-chairs the NFDI Working Group on Overall Architecture (Politze et al., 2025), has been a member of the task force defining the NFDI EOSC Node since its inception and aligns its own architectural developments accordingly. Legal and Regulatory Developments: The evolving legal landscape around the use of (research) data and AI has had a direct impact on the work of Text+. Key developments include the EU AI Act (European Parliament, 2024), the EU Data Act (European Commission, 2024), and the proposed German Research Data Act (Bundesministerium für Bildung und Forschung, 2024). These frameworks influence how data can be shared, processed, and reused – not exclusively, but especially in the context of AI. Text+ has taken an active role in shaping these discussions through its involvement in the NFDI section Ethical, Legal, and Social Aspects. Text+ experts have contributed to several key policy statements and consultations, including: Opinion on the EU Data Act Proposal (NFDI e.V., 2022); Stellungnahme zum Daten-Governance-Gesetz (Opinion on the Data Governance Act) (NFDI e.V., 2023a); Stellungnahme zur öffentlichen Konsultation zum Forschungsdatengesetz (Response to the consultation on the German Research Data Act) (NFDI e.V., 2023b).

5. Work programme

Table 5.1 Overview of the task areas

Task Area	Measures	Co-Spokespersons
Library and Archive Holdings as Research Data	(1) Federated Research Data Infrastructure (2) Interfaces between Data from Data Domains (3) Enabling Text and Language Data for AI (4) Implicit and Explicit Structures of Language (5) Community Activities	Philippe Genêt (DNB) Elke Teich (UdS)
Language Corpora and Elicited Data		Andreas Witt (IDS) Angelika Zirker (EKUT)
Lexical Resources		Hanna Fischer (UMR) Alexander Geyken (BBAW)
Editions		Andrea Rapp (TUDa) Andreas Speer (AWK-NRW)
Infrastructure / Operations	(1) Federated Research Data Infrastructure (2) Service Integration with Data Domains (3) Platform and Core Services (4) Coherence	Vivien Petras (HUB) Regine Stein (VZG) Philipp Wieder (UGOE/SUB)
Administration	(1) Financial Administration (2) Flexible Funds (3) Project Office (4) Board Office (5) Communication Office (6) NFDI Cooperation	Philipp Wieder (UGOE/SUB) Andreas Witt (IDS)

In phase II, the structure of the Text+ work programme builds upon the proven framework established during phase I. With now four data domain-specific task areas and two transversal task areas, the work programme reflects both the thematic focus of Text+ in phase II (see also section 2.1) and the operational backbone required to support a sustainable research data infrastructure.

All Text+ task areas are relevant to other NFDI consortia in many ways beyond Text+ and contribute to cross-cutting topics, so only examples are mentioned here: TA1 Library and Archive Holdings as Research Data focuses on the processing and enrichment of unstructured text, such as for AI applications. This is a challenge that also arises in other subject areas. TA2 Language Corpora and Elicited Data develops corpora for the use of large language models, for example; this technology is also being researched in other consortia. TA3 Lexical Resources is strongly linked to terminologies and controlled vocabularies in general, which are widely used across subject areas to support FAIR data. TA4 Editions is relevant across the humanities, for example in musicology and thus in NFDI4Culture. TA5 Infrastructure/Operations is by definition the interface to other NFDI consortia for the harmonisation of cross-cutting services and the connection of basic services. TA6 Administration facilitates as governance unit the inter-consortial exchange, for example, in the NFDI management circle and coordinates activities within the NFDI Association.

5.0.1 Structure and role of the cross-cutting measures of the data domains

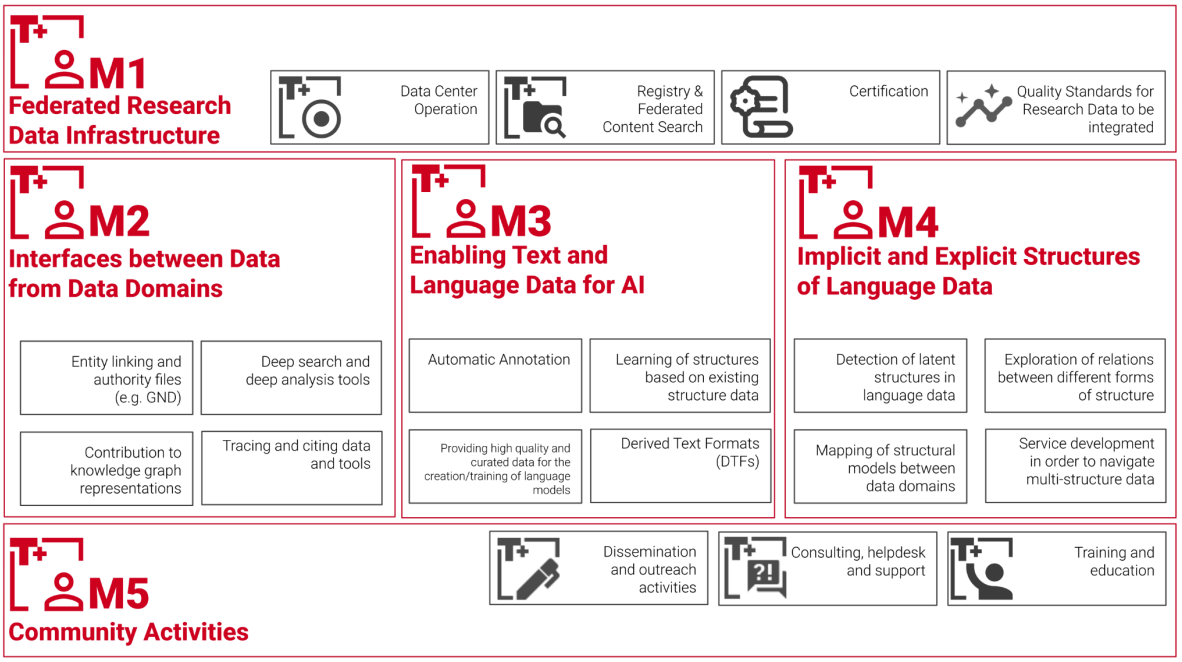


Figure 5.1: Cross-cutting measures of TA1–4 with associated activities

As in phase I, the task areas representing the data domains (TA1–TA4) collaborate through a coordinated set of cross-cutting measures. In phase II these are M1: Federated Research Data Infrastructure; M2: Interfaces between Data from Data Domains; M3: Enabling Text and Language Data for AI; M4: Implicit and Explicit Structures of Language; and M5: Community Activities.

These measures build directly on earlier efforts related to measures from phase I. By refining the measure structure and deepening its integration across task areas, phase II aims to ensure even greater conceptual alignment, operational coherence, and responsiveness to community and technological developments. These measures also serve as core instruments to implement the strategic themes (ST1–4) and project-wide objectives (O1–9) defined in sections 2.1 and 2.2 and are complemented by specific contributions within each task area. See Table 5.1 for the strategic alignment of the measures with the project-wide strategic themes and objectives.

Table 5.2 Strategic alignment of the measures across the data domains

Measure	Strategic Themes	Objectives
M1: Federated Research Data Infrastructure	ST3	O1, O2, O5, O6, O7, O8, O9
M2: Interfaces between Data from Data Domains	ST1, ST2	O2, O4, O7
M3: Enabling Text and Language data for AI	ST1, ST2	O2, O4, O6, O8
M4: Implicit and Explicit Structures of Language	ST2, ST3	O3, O6
M5: Community Activities	ST3	O1, O5, O6, O7, O8, O9
Note: not applicable for TA Admin. Alignment for TA IO is explained in the respective chapter.		

The following descriptions present the cross-cutting aspects of each measure, including their specific objectives, implementation activities, success indicators, deliverables, and risks. Task area-specific extensions and resource allocations are detailed in 5.1 to 5.4.

Measure 1: Federated research data infrastructure

Cross-TA responsibility: Philipp Wieder (TA5) and Andreas Witt (TA2)

Contributors: All partners from TA1–TA4 in the respective measure 1. Please note that M1 is **supported completely by in-kind contributions of the Text+ Centers** (see 5.1 - 5.4 for details regarding the respective amounts), and no additional funding is requested.

Objective: M1 aims to maintain and progress the sustainable, distributed Text+ Data Space that provides coherent access to research data from all four data domains of Text+. The actual development of the services to be integrated will be carried out by a measure of the same name in the task area “Infrastructure/Operations” (TA5 M1; see section 5.5).

Key activities: Key activities focus on the continuous operation of the Text+ Centers, their integration into an evolving Text+ Data Space, and their connectivity to the NFDI as a whole. In detail, this includes:

- Ensuring the long-term sustainability of the Text+ Centers.
- Defining and implementing content-related quality standards for data.
- Offering services for ingesting new data sets into the Text+ Data Space.
- Provisioning of standardized, high-quality metadata in particular through the Text+ Registry.
- Operating and updating of a dedicated FCS endpoint.
- Registering and maintaining up to date entries in re3data.
- Integrating Text+ core services (TA5 M3) as well as basic services from the growing portfolio of the NFDI (like IAM4NFDI or PID4NFDI) into Text+ Centers.

Relevance to goals of the NFDI: M1 directly contributes to the NFDI's mission to provide high-quality, interoperable, and sustainable data services. It also strengthens the connection between humanities scholars and research infrastructure.

Success indicators: Number of resources recorded in the Text+ Registry; number of metadata records provided by participating centers; number of PIDs assigned to data sets; number of research data sets ingested within a defined timeframe, number of queries executed via the FCS; existence of valid certifications for Text+ Centers; accuracy and currency of center entries in re3data.

Table 5.3 M1 joint deliverables (synchronised yearly with NFDI reporting)

#	Month	Deliverable	Notes
D0.1.1	12, 24, 36, 48, 60	Report on the status and contributions of the Text+ Centers as main contributors to the Text+ Data Space, e.g. listing in discipline-agnostic directories of repositories (e.g. re3data.org), certification status, data sets, number of users, publications	Contr. to appendix 5 fields 4001 & 4002, 4004–4008 and 5001–5109 in the category "storage". Corr. to appendix 5, fields 3001 & 3002, 1001 & 1002. See also objective O5

Note: Deliverables and milestones are labeled in a consistent scheme: each identifier begins with a "D" (for deliverables) or "M" (for milestones), followed by the task area number, the measure number, and a sequential number within that measure. Joint deliverables or milestones that apply across all task areas use "0" in place of the task area number.

Table 5.4 M1 risks of implementation (TA-specific risks are listed in TA1–4)

Risk identified	Risk prevention and mitigation
Bottlenecks in metadata integration and registry workflows	Promote automation and scalability in metadata workflows; provide shared documentation and technical support; explore semi-automated ingestion pipelines to reduce manual workload.
Limited scalability of the Text+ Registry and FCS infrastructure	Collaborate with TA5 to improve backend performance and interface design; prioritise development of scalable APIs and metadata harvesting protocols.
Legal and institutional constraints	Provide consultation and legal review services; ensure IAM integration and compliance with institutional and national frameworks.

Fragmentation due to institutional autonomy	Maintain a federated but harmonized infrastructure model; align technical standards and service interfaces across centers while respecting autonomy.

Measure 2: Interfaces between data from data domains

Cross-TA responsibility: Angelika Zirker (TA2) and Alexander Geyken (TA3)

Contributors: All partners from TA1–4 in the respective Measure 2

Objective: M2 will establish interfaces between research data from various domains, sources, and disciplinary backgrounds within the Text+ Data Space. It aims at defining interfaces not only within the data domains themselves but also between the services and tools used for research and analysis. Through this, Text+ enables deep search and data-driven insights based on entity linking, authority files, and knowledge graphs.

Key activities: Key activities focus on establishing interoperable interfaces across metadata, object data, and analytical tools to enable cross-domain integration within the Text+ Data Space. This includes:

- Integrating persistent identifiers and metadata standards (e.g. ISO 24619) to support users with deep search as well as tracing and citation of data.
- Designing and implementing interfaces linking entities across data sets via authority files, e.g. by using the GND.
- Building a knowledge graph-based infrastructure for semantic interlinking of resources.
- Connecting tools and services tailored to specific data types within the federated infrastructure (see also M1).
- Promoting reuse through entity linking, semantic annotation, and alignment with FAIR principles.

Relevance to goals of the NFDI: M2 is essential to the broader NFDI goal of fostering robust research data management and bridging disciplinary divides. By establishing both technical and conceptual connections between data sets, M2 supports sustainable reuse, citation, and collaborative research practices.

Success indicators: Number of entity reference links in (meta)data; number of shared entities across data sets; number of research outputs based on data from multiple Text+ Centers; integration of tools and services; referencing data sets in deep analysis tools and services; outreach: participation in workshops; availability of ground truth data sets.

Table 5.5 M2 joint deliverables

#	Month	Deliverable	Notes
D0.2.1	12	Report on representing Text+ metadata as knowledge graphs	./.
D0.2.2	24	Report on including entity links in metadata representations	./.

D0.2.3	48	Provision of a technical interface for querying linked data, including: interface implementation, documentation, example queries	./.
D0.2.4	48	Workshop for educational purposes to present possible data uses from the federated infrastructure to deep analysis tools and services	Contributes to Datasheet field 1008.
D0.2.5	54	Monitoring platform to track the number and usage of entity links in metadata and object data	./.

Table 5.6 M2 risks of implementation (TA-specific risks are listed in TA1–TA4)

Risk identified	Risk prevention and mitigation
Lack of shared understanding of interface requirements across domains	Organise cross-domain regular working group (WG) meetings early in the implementation phase; develop shared documentation and use cases to align expectations
Inconsistent use of authority files and entity linking practices	Provide training and guidelines in the joint WG; promote use of shared vocabularies (e.g., GND, Wikidata); implement validation tools for entity linking
Technical complexity of integrating heterogeneous metadata and object data	WG develops modular, standards-based interfaces; reuse existing tools and protocols (e.g., CMDI convertors to JSON-LD and other convertors); collaborate closely with TA5
Limited performance of knowledge graph infrastructure	Joint WG works on pilot with representative data sets; monitor performance; plan for incremental scaling and optimization of query endpoints
Low adoption of interfaces or linked data by researchers	WG provides use cases and demonstrators; organise training and outreach events; integrate interfaces into workflows and tools

Measure 3: Enabling text and language data for AI

Cross-TA responsibility: Philippe Genêt (TA1) and Vivien Petras (TA5)

Contributors: All partners from TA1–4 in the respective Measure 3.

Objective: M3 focuses on preparing, enriching, and utilising text and language data for use in artificial intelligence (AI), particularly machine learning (ML) and large language models (LLMs). The goal is to enable responsible, high-quality AI applications in the humanities and text- and language-based research, and to keep up with the rapid pace of innovation.

Key activities: M3 addresses the following key areas:

- Provisioning of AI-based tools for automatic annotation of text and language data, such as annotations on multiple tiers of spoken data.
- Learning of structures based on existing structure data like learning document hierarchies, metadata patterns, or typographic structures from legacy collections.
- Provisioning of high quality and curated data for the creation and training of language models (also serving subjects outside Text+ like law history and medical dissertations).
- Creation and provisioning of Derived Text Formats (DTFs) (Schöch et al., 2016) that allow the reuse of rights-protected text and language data in machine learning and other distant analysis contexts.

- Integrating (semi-)automatic data curation, code revision, and documentation. With the productive use of AI in co-piloting, code assistance, the cost of reviewing and refactoring (e.g. web-based scholarly editions) can be significantly reduced.
- Another key challenge to be addressed is bias, both in the underlying data and in AI models. M3 will develop methods to evaluate and mitigate bias by providing representative data sets and metadata, developing transparent annotation pipelines, and fostering inclusive AI development practices.

Relevance to goals of the NFDI: M3 supports the overarching goals of the NFDI by promoting sustainable, legal, and effective reuse of research data, for example in AI contexts. It fosters collaboration across disciplines and institutions, strengthens methodological innovation, and ensures that language data is FAIR for ML applications.

Success indicators: Number of automatically annotated corpora made available; number of structured data sets made available for LLM training; uptake and reuse of AI-ready data sets by the research community; number and diversity of DTFs produced and documented; availability of software and data sets curated with AI-assistance; availability of bias assessment tools or documentation for training data sets and models; participation of external stakeholders in AI-focused workshops and training events.

Table 5.7 M3 joint deliverables

#	Month	Deliverable
D0.3.1	12	Guidelines for automatic annotation workflows and quality control
D0.3.2	18	Use cases exemplifying AI applications of Text+ resources
D0.3.3	30	Technical documentation for DTFs and conversion pipelines
D0.3.4	48	Bias documentation templates and evaluation frameworks
D0.3.5	18, 36, 54	Workshops on best practices for using text and language data in AI

Table 5.8 M3 risks of implementation (TA-specific risks are listed in TA1–TA4)

Risk identified	Risk prevention and mitigation
Legal and ethical constraints on using personal or copyrighted data for AI training	Joint WG with regular meetings to develop and promote Derived Text Formats (DTFs); implement anonymisation workflows; ensure legal review and compliance with copyright and data protection laws.
Lack of interoperability between AI-ready data and existing infrastructure	Joint WG to develop standard APIs and conversion pipelines; ensure compatibility with federated access services and metadata standards.
Fragmentation in tool development or duplication of effort	Joint WG with regular meetings to monitor emerging technologies; maintain modular, updatable workflows; align with NFDI-wide AI initiatives; maintain a shared list of tools and models; encourage reuse and modularity.

Unclear attribution for AI development and data contributions	Implement and develop practices and guidelines on extending rich metadata for provenance and authorship, including procedures to make AI-based or assisted workflows transparent; adopt citation standards for data sets and models.

Measure 4: Implicit and explicit structures of language

Cross-TA responsibility: Andrea Rapp (TA4), Regine Stein (TA5), and Elke Teich (TA1)

Contributors: All partners from TA1–4 in the respective Measure 4

Objective: M4 addresses the heterogeneous and layered structural representations found in language- and text-based research data. It systematically analyses the relationships between implicit and explicit structures across Text+ domains to provide infrastructure that enables researchers to navigate across data types and levels of formality, without enforcing reductive or uniform frameworks.

Key activities: M4 explores the relationships between the structures of text and language data, and how these structures manifest in the specific data types found in the Text+ task areas. This includes: Library holdings represent the relationship between ePub and TEI formats. OCR data, including pagination, layout, text topography, and metadata, encodes structural interpretations of printed books. Corpora examines how syntactic treebanks relate to flatly segmented corpora and how various annotation layers – such as morphological, syntactic, and discourse – coexist or conflict. In literary data, narrative or rhetorical structures in literary texts become visible and traceable through annotations or computational tools that support literary and interpretive analysis. Metadata explores how register structures and catalogue metadata interact with access structures in language-based research data. Finally, editions reflect the implicit knowledge, such as methodological decisions, assumptions, and biases, that influence how a source is represented but are often not made transparent to users.

M4 will, inter alia, work on mapping of structural models across data types, exploring how “unstructured” data (e.g. plain text) becomes structured through scholarly or technical intervention. M4 will also focus on the documentation of implicit structures.

Success indicators: Number of structural models across data types in Text+; number of data sets in the federated infrastructure of tools or services enabling access to and visualisation of multi-structured data; availability of crosswalks or mappings between structural formats; case studies exemplifying transitions from implicit to explicit structure; best practices for representing implicit structures or uncertainty; documentation describing the structural logic of Text+ data types; integration of structuring models into annotation workflows and data ingestion; recognition of structuring diversity in metadata schemas and user interfaces.

Relevance to NFDI goals: Implicit and explicit structures are inherent to all forms of information, e.g. from data tables to programming languages. Language data, with its

complex, multi-layered, and often ambiguous structure, presents one of the most challenging data types to model. Developing approaches to structure, annotate, and document, or even building knowledge graphs from data structures, offers conceptual and technical models relevant to all other NFDI consortia.

Table 5.9 M4 joint deliverables

#	Month	Deliverable
D0.4.1	12	Documentation of structural models within and across all major Text+ data types (e.g. editions, corpora, treebanks, metadata collections)
D0.4.2	18	Cross-domain comparative reports on structural conventions (e.g. comparison of narrative structure, annotation logic, or layout representation)
D0.4.3	30	Mappings and transformation guidelines between structuring models
D0.4.4	36	Annotated data sets and examples showing implicit / explicit structuring processes
D0.4.5	54	Shared technical components and interfaces that support navigation and querying across multiple structures
D0.4.6	12, 30, 54	White papers reflecting epistemological implications of “structurification” across domains.

Table 5.10 M4 risks of implementation (TA-specific risks are listed in TA1–TA4)

Risk identified	Risk prevention and mitigation
Conceptual ambiguity in defining implicit, explicit, and latent structures	Develop shared terminology and conceptual frameworks; organise cross-domain WG to align understanding and use of structural categories.
Incompatibility between structural models across data types and domains	Create mappings and transformation guidelines; promote modular annotation frameworks; support coexistence of multiple structural views in infrastructure.
Technical challenges in supporting multi-structure navigation and querying	Collaborate with TA5 on interface design; pilot technical components with representative data sets; ensure extensibility and scalability of tools.
Lack of standardized documentation for structural logic and transitions	Produce structured documentation templates; encourage contributions from all domains; integrate documentation into ingestion and annotation workflows.
Limited adoption of multi-structure tools and practices by researchers	Provide training and demonstrators; publish best-practice examples and case studies; integrate tools into familiar research environments.
Oversimplifying structural diversity in metadata or interfaces	Design infrastructure to support parallel structural representations; avoid enforcing reductive schemas; promote epistemological awareness in design decisions.

Measure 5: Community activities

Cross-TA responsibility: Hanna Fischer (TA3) and Andreas Speer (TA4)

Contributors: All partners from TA1–4 in the respective Measure 5.

Objective: M5 focuses on engaging with and connecting to research communities by communicating and providing Text+ services, knowledge, and support relating to text- and language-related research data (management). Text+ will consolidate its community approach by streamlining the disseminating formats, educational outputs, and its consulting

workflows, to comply with NFDI-wide initiatives. M5 will also connect to EU-wide structures, such as EOSC, e.g. via the NFDI EOSC National Node.

Key activities: By addressing the needs of researchers, M5 aims to encourage the further adoption of the infrastructure and resources developed in, distributed by, and provided through Text+. This will help to ensure continuous contribution to a sustainable cultural change in data practices. To implement this, M5 includes the following key activities:

- Implementing dissemination and outreach activities.
- Creating training and education output.
- Providing RDM consulting service, including a helpdesk for central services of Text+ and those contributed by partners.

Thus, Text+ will expand its focus on networking and cooperating with further communities (e.g. legal studies and various philologies) and their stakeholders providing services to the scholarly community of Text+. Among these communities are learned societies, Specialized Information Services (FID), and other NFDI consortia providing services for the humanities. A core principle of M5 is user orientation: researchers and data providers are not just recipients of services, they actively shape the requirements and future directions of the infrastructure. This involves constant evaluation of communication tools such as the Text+ Portal, Text+ Blog, and other public-facing resources, to customize them so they meet the needs of target audiences. M5 emphasizes the value of the services, data products, and expertise developed and provided in Text+, ensuring they are accessible, understandable, and useful.

Relevance to goals of the NFDI: M5 supports the NFDI's commitment to community-driven research data infrastructures. By building bridges between technical infrastructure and everyday research practice, it fosters capacity building, encourages reuse, and guarantees that NFDI investments benefit a broad and diverse spectrum of disciplines and institutions.

Success indicators: Number of training events, workshops, and educational materials; participation levels in outreach activities, especially among researchers outside the consortium; number of consultations and use cases supported; usage statistics and satisfaction feedback for the helpdesk; number of (successfully closed) helpdesk tickets; improved usability and accessibility of public-facing platforms, such as the Text+ Portal; collaborations with subject-specific learned societies, research initiatives, and library-based services; uptake of Text+ tools and data resources among new user communities; community feedback loops that influence infrastructure development.

Table 5.11 M5 joint deliverables

#	Month	Deliverable	Notes
M0.5.1	36	Helpdesk infrastructure and knowledge base integration into NFDI wide structures.	
D0.5.2	12, 24, 36, 48, 60	Annual report to CCs on outreach campaigns (incl. blog posts), publications, educational materials (incl. guides, tutorials), events (incl. workshops, hackathons, stakeholder meetings), and documented consultations.	Synchronised yearly with NFDI reporting; corr to appendix 5, fields 1007-1009, 1021, 2001

Table 5.12 M5 risks of implementation (TA-specific risks are listed in TA1–TA4)

Risk identified	Risk prevention and mitigation
Limited engagement from new or underrepresented communities	Conduct targeted outreach through stakeholder mapping; collaborate with SIS, learned societies, and partner consortia to reach relevant groups.
Low uptake of training and educational resources	Align training formats with community needs; promote through established networks; make materials accessible, modular, and reusable.
Inconsistent consulting quality or documentation across TAs	Standardize consulting workflows and documentation; align with NFDI-wide helpdesk structures; provide internal training for consultants.
Overload or fragmentation of helpdesk services	Integrate helpdesk into the federated NFDI helpdesk network; streamline internal workflows; clear routing of inquiries.
Difficulty in demonstrating added value of Text+ services to users	Use persona-based design and user feedback to improve usability; highlight concrete use cases and success stories in outreach.
Duplication of outreach or traini efforts across domains	Coordinate dissemination and training activities across task areas; maintain a shared calendar and cloud-based directory of materials.
Rapid evolution of RDM and AI-related training needs	Monitor emerging trends; update training content regularly; collaborate with NFDI Section Education and Training and EOSC initiatives.

5.1 TA1: Library and Archive Holdings as Research Data

Co-leads: DNB, UdS

Contributors: DIE, DLA, HAdW, IDS, JMU, KSW, LMU, MWS, SAW, UGOE/SUB, TUDa, UHH, UniTR, UzK

Task Area 1 (TA1) focuses on the largest, most heterogeneous, and least structured domain of data. TA1 holdings include digital texts and audio-visual documents ranging from texts (e.g. novels, essays, newspapers), to audio files (e.g. radio programmes, speeches, podcasts), videos from different sources (e.g. television, YouTube content, scientific field recordings), and multimodal objects (e.g. comics, exhibition catalogues, audiobooks). A common feature of these resources is that, beyond transcribed text and metadata, the data typically does not include explicit structural information (e.g. token, sentence, or paragraph

mark-up). Their low level of internal structural information is the greatest potential but also the hardest challenge for TA1 holdings. The potential lies in making a vast set of materials available for scientific engagement with language- and text-based data across a wide range of disciplines (see strategic theme 1, objective 5). In Text+, TA1 holdings serve as the starting material for more structured forms of research data (such as corpora, editions, or lexical resources), which, in turn, flow back as the result of scientific investigation (e.g. as dissertations, digital scholarly editions, dictionaries). The main challenges involved in empowering TA1 holdings to become research data are, first, that data must be made *findable, accessible, interoperable*, and *reusable* following the *FAIR* principles (see Grand Challenge 1). This is achieved through their integration into the Text+ Data Space as well as through the use of established data standards, versatile formats, and enrichment with links to authority data. Second, the initial technical steps towards “more structure” must be made available to researchers so as to enable them to turn “raw data” into data that can serve as a basis for corpora, editions, lexical resources, and other kinds of research data (e.g. converting plaintext into common input formats for linguistic annotation). For this purpose, Text+ offers dedicated tools – such as the NLP pipeline *MONAPipe* (Dönicke et al., 2022) from UGOE/SUB and the web services for automatic transcription and speech processing – but also provides consultation and support in selecting and, if desirable, adapting other tools that are commonly used in the community (e.g. CQP, KorAP, SpaCY, Stanza, Emu/R).

TA1 holdings are to be made as accessible as possible for various disciplines, including those that do not belong to the core community of Text+, e.g. (oral) history, law, medicine, sociology, economics (see objective 1). To this end, thematic or disciplinary collections across the holdings of different Text+ Centers will be curated and made available, and individual researchers will be enabled to compile virtual collections. Material from TA1 holdings will also be offered for training language models. Finally, the work programme of TA1 also addresses the continuous exploration of resources that have previously been inaccessible to research. This includes (1) integrating third-party holdings into the Text+ Data Space and continuing the exploration of holdings of TA1 Centers, particularly into the FCS, (2) providing copyrighted material in derived text formats (DTFs), i.e. formats that do not allow to reproduce the original text but can still answer specific research questions (Schöch et al., 2016), (3) offering assessment by legal experts on copyright and licensing laws, and/or privacy and data protection, (4) providing support regarding ethical considerations (e.g. concerning the training and use of language models) for both the Text+ partner institutions and the scientific community more widely, and (5) widening the focus for bibliographic data which is at the core of TA1 holdings.

TA1 complements TA2 Corpora and Elicited Data where the focus is on subsequent stages of data processing, such as dedicated linguistic annotation.

TA1 M1: Federated research data infrastructure

Co-leads: UdS, DNB

Contributors: DIE, IDS, LMU, MWS, SAW, UGOE/SUB, UHH, UniTR, UzK

The Text+ Centers in TA1 will contribute to the further development, maintenance, and expansion of the federated research data infrastructure set up during the first funding phase. TA1 institutions will update and continuously add data sets to the Text+ Data Space, including in-copyright material that will be made available in DTF (e.g. by DNB, SAW and UniTR), extensive library catalogues with access to e-publications and full texts (e.g. by MWS), further digitized sections of VD17 and VD18 held by UGOE/SUB (also by encouraging more libraries outside the consortium to integrate their data in this particular regard), multilingual written and spoken language collections (e.g. by UdS and UHH), phonetically relevant spoken language collections (e.g. by LMU), and audiovisual resources from typologically diverse languages (e.g. by UzK).

TA1 will also continue to offer the established Text+ services such as the data depositing service for research data from diverse communities (e.g. in the TextGrid Repository run by UGOE/SUB, the CLARIN-B centre of the UdS and the spoken language repositories of UHH, LMU and UzK), RDM-consultancy, (meta)data curation, and long-term archiving.

Software services will be continuously operated, evaluated, and developed, including the spaCy-based (ExplosionAI GmbH, 2025) NLP pipeline MONAPipe with its modularized integration of community-relevant components alongside an infrastructure designed in accordance with research software engineering principles, the digital publishing service perspectivia.net (run by MWS), as well as transcription tools for digitized texts (e.g. OCR and multimodal AI-based approaches to generate plain text from digitized facsimiles). TA1 will test and develop processing pipelines to get from “raw data” to a corpus with tools commonly used in the community, with a focus on languages other than German as well as multimodal data (e.g. data preparation and publication processes, including the implementation of a graphical web interface for the corpus services of UHH). Above that, TA1 repositories will pursue certification wherever the process is not accomplished yet or renew certificates. DNB will support this with expertise especially in the case of Nestor certificate. CLARIN and DARIAH partners in TA1 (e.g. UGOE/SUB, MWS, UdS, SAW) will bring in expertise on CoreTrustSeal certification.

Similarly, the *Fluffy TextGrid Publication Workflow* developed by the UGOE/SUB, GWDG, and TUDD as a new import and publication option via the TextGrid Repository will be continued and rolled out more widely (see also section 5.5.1). Implemented as a Jupyter

notebook, *Fluffy TextGrid Publication Workflow* offers a user-friendly GUI for easy import and supports high-quality metadata through standards (e.g. GND entities, basic classification, and ISO code for languages).

In-kind contributions by partners (in PM): DIE (30), DNB (24), UdS (33), UGOE/SUB (30), IDS (30), MWS (30), UniTR (15), UHH (30), LMU (30), SAW (6), UzK (6)

All achievements of TA1 in M1 correspond to the joint deliverables of M1 in section 5.0.

TA1 M2: Interfaces between data from data domains

Lead: DNB

Contributors: IDS, MWS, SAW, UGOE/SUB, UdS, UHH, UzK

While structural features are inherent in TA1 data that have not yet been made explicit, its resources typically have rich metadata that can serve as an interface, e.g. for search tools such as the Text+ Registry. TA1 contributes to M2 by providing tools, methods, and knowledge to improve and enrich metadata for resources across all data domains. The key areas of M2 will be addressed as follows.

Activity 1 – Entity linking and authority files: A special focus lies on the use of authority files such as GND (*Gemeinsame Normdatei*) – provided by DNB – to link named entities to unique identifiers for unambiguous identification. Consistent entity linking on metadata level as well as within resources will permit sophisticated means of exploration and help discover previously undetected connections between resources, which may lead to new research questions. TA1 will continue to work on providing entity linking tools, e.g. in the NLP pipeline MONAPipe, and also aims to enhance the searchability and interconnectivity of resources within the Text+ Data Space by developing tools and technologies for automatic tagging, keywording, and linking of resources to authority file subject headings IDs as controlled vocabulary, as well as for building interfaces across different modalities and language data tiers.

Activity 2 – Deep search and deep analysis tools: The FCS will be adapted to the special requirements of TA1. The aim is a user-friendly search in – partly restricted – collections for further use (e.g. LLM training). The selection of resources for their processing in external working environments (such as HPC environments or NLP pipelines) will be supported by search functionalities (e.g. based on entities and allowing searching in DTFs). Within Text+, the corpus analysis platform KorAP serves as a reference implementation for deep search solutions, particularly for holdings protected by copyright. With comprehensive documentation, installation routines, and testing frameworks, KorAP offers an off-the-shelf solution that supports FCS and can be readily adopted by other centers. While its use is optional, KorAP provides a flexible option for those seeking a robust analysis platform tailored to their data needs.

Activity 3 – Contribution to knowledge graph representations: In order to connect their holdings to Knowledge Graphs (e.g. in the wider NFDI context), TA1 will implement Knowledge Graph endpoints to increase the degree of interoperability. Above that, linked open data (LOD) and resource description framework (RDF) (W3C, n.d.b) representations of TA1's resources will ensure connectivity to various data infrastructures in the NFDI and beyond. The Annotation Store, developed by UGOE/SUB (see 1.1.4), will provide a SPARQL (W3C, n.d.c) endpoint to meta- and object data.

Activity 4 – Tracing and citing data and tools: With the help of long-standing library-specific experience in metadata enrichment, TA1 will help to further increase the findability of resources in the Text+ Data Space and beyond. By using established standards, terminologies, and ontologies, TA1 contributes to the interoperability of Text+ resources (data *and* tools), especially regarding interfaces to other infrastructures such as the NFDI, OPERAS, SSHOC, and EOSC. Text+ started these activities with the integration of the TextGrid Repository in the EOSC Resource Hub and the integration of the DNB catalogue data on record level for the selected SSH disciplines as well as the catalogue data of perspectivia.net in GoTriple (OPERAS). TA1 will continue this path by monitoring the development within the EOSC regarding interoperability. Furthermore, the Text+ Registry will be used for drawing a landscape of TA1 holdings including a defined categorization of usability of the provided data for broader use cases. Finally, the work plan also comprises prioritization and exemplary implementation of metadata for literary studies (Calvo Tello et al., 2024a).

In-kind contribution by partners (in PM): DNB (7), IDS (30), UdS (12), UGOE/SUB (6)

Requested funding (in PM): DNB (20), IDS (8,87), MWS (12), SAW (12), UGOE/SUB (10), UdS (13), UHH (10), UzK (3)

Table 5.13 Deliverables and milestones for TA1 M2

#	Month	Deliverable
D1.2.1	18	Report on TA1 holdings and their usability for defined use cases
M1.2.2	24	Reference implementation of KorAP completed
D1.2.3	30	Report on the exemplary integration of dedicated metadata for literary studies
M1.2.4	36	Integration of FCS adaptations to TA1 needs
D1.2.5	42	Evaluation report for provided entity linking and tagging tools

TA1 M3: Enabling text and language data for AI

Lead: UdS

Contributors: DNB, LMU, SAW, UGOE/SUB, UHH, UniTR, UzK

TA1 addresses both text and spoken modalities to make available on a large scale material from TA1 holdings as well as spoken language records, and to provide reusable, robust

workflows for curating and annotating new material in the future. Specifically, TA1 addresses the key areas defined for M3 as follows.

Activity 1 – Automatic annotation: The distinctive contribution of TA1 to automatic annotation with LLMs is generation of relevant metadata. While bibliographic data (author, time and place of publication, edition etc.) are clearly essential for searching, the content keywords assigned through libraries are often too coarse or insufficient from the perspective of text and language scholars. Using prompting and few-shot methods, texts can be better assessed with LLMs in terms of text type (e.g. factual report vs. literary essay), register (field/topic, tenor/attitude, spoken vs. written mode of discourse), or (literary) genres (e.g. science fiction vs. romance). Recent studies have shown that LLMs perform quite well at text classification tasks (Henriksson et al., 2024; Alizadeh et al., 2024; Pichler et al., 2025).

TA1 plans to develop specific scripts and to test various classification schemas from linguistic as well as literary perspectives, e.g. register theory, *Textsortenlinguistik* and literary genre theory, as well as to compare various models (e.g. GPT, LLaMa). While TA1 will focus on German and English, other languages will be considered as well, depending on community interests. Once texts have been annotated with relevant metadata, they can be used to fine-tune existing language models or train new models, or simply to build a corpus.

Activity 2 – Providing high quality and curated data for the creation/training of language models: TA1 will focus on freely available material (e.g. dissertations, scientific journal articles, annotated data from digital editions, multilingual corpora) for LLM training and fine-tuning, again including text and spoken language data. The goals of TA1 are: (a) to provide text and spoken language materials on a large scale that can be used for LLM training etc., (b) to openly share task-specific LLM-based workflows and processing pipelines, and (c), for the spoken language community, to develop a shared community model tailored to their specific needs in the form of a customizable and reusable resource that reflects the linguistic diversity and methodological requirements of the field.

Activity 3 – Derived Text Formats (DTFs): For use of copyright-protected material for computational processing and subsequent analysis, TA1 has developed DTFs. To the extent that language model training (or fine-tuning) based on DTFs delivers satisfactory results, they are an excellent opportunity to make available otherwise inaccessible material. TA1 plans to carry out several studies to showcase the application of DTFs on various data sets (languages, types of publications). Apart from providing selected materials in DTFs for LLM training or other purposes, TA1 will integrate DTFs into MONAPipe as an additional module. DTFs can then be published in the TextGrid Repository as a new data type.

Activity 4 – Addressing bias: At any point, TA1 critically reflects the use of LLMs compared to established methods (e.g. for annotation) as well as evaluates LLM output in terms of quality, including potential bias (objective 4; Gallegos et al., 2024). Furthermore, TA1 will work towards a standard schema to describe biases inherent in language data (e.g. as a special kind of metadata that comes with a resource) in order to help scholars choose appropriate training material for their respective needs.

All partners participating in M3 provide advice regarding the application of scripts for LLM use in selected tasks (e.g. transcription, classification) as well as on ethical and legal issues concerning the use of LLMs and AI.

In-kind contributions by partners (in PM): DNB (7), UdS (12), UGOE/SUB (6)

Requested funding (in PM): DNB (30), LMU (3), SAW (6), UGOE/SUB (10), UdS (58), UHH (30), UniTR (18), UzK (6)

Table 5.14 Deliverables and milestones for TA1 M3

#	Month	Deliverable
D1.3.1	18	Pilot study on DTF use (+ publication)
M1.3.2	24, 36, 42	Publication of material in DTFs, transcribed material, enriched metadata using LLMs
D1.3.3	30	Pilot studies transcription and text classification (+ publication)

TA1 M4: Implicit and explicit structures of language data

Lead: UGOE/SUB

Contributors: DNB, LMU, UdS, UHH

Typical structures in TA1 holdings include, e.g. pages, chapters, paragraphs, lines, words, and characters in text resources, whereas audio material can feature, e.g. pauses, hesitations, intonation, or laughter. TA1 will provide tools, documentation, and support to analyze these kinds of structural information in resources of any modality and make them explicit – either in their metadata or as annotation layers. Deployed annotations can be further integrated in the FCS, e.g. via SPARQL endpoints, in order to enhance search and retrieval results. A special focus lies on exploring the relations between ePub and TEI representations, on structural information inherent in OCR data, and on the structuring role of metadata, e.g. in library and archive catalogues.

These and other kinds of structural information that may be relevant to the various Text+ communities will be assessed by surveys and integrated in annotation tools such as the NLP pipeline *MONAPipe* that integrates and deploys models for the automatic annotation of text and language data and is part of UGOE/SUB's infrastructure for the creation, publishing, and presentation of structured annotations. The existing MONAPipe will be constantly extended by relevant components that are developed within the research community – among these are models based on classical machine learning, deep learning,

and LLM-driven approaches (generative AI). In addition to the local usage, computing for complex neural models will be integrated in the high performance cluster of the GWDG alongside with dedicated APIs. To ensure the interoperability of the generated annotations, existing export formats (like CoNLL-U) will be expanded with further options for formats used in the community (such as TEI). To enable even more direct user interaction, TA1 will experiment with the integration of front-end interfaces (e.g. no-code/low-code solutions). In the *Web Annotation Protocol Compliant Annotation Store (FE / WAPCAS, 2025)*, stand-off annotations generated by MONAPipe can be stored, published, and delivered following the Web Annotation Data Model, a W3C recommendation since 2017. The Annotation Store resolves the relationship between the annotations and textual data and also offers public APIs (e.g. to deliver annotations to the FCS). In order to present annotations, the *TIDO Viewer* offers components for displaying such annotations based on the W3C specification. Beyond that, TA1 supports other Text+ Centers in the integration of further annotations or textual relations – not limited to object data in text form but also allowing structured representations between all kinds of resources that can be hosted on the web, e.g. object data in audio, video, and other binary formats.

In-kind contributions by partners (in PM): DNB (7), UdS (12), UGOE/SUB (9)

Requested funding (in PM): DNB (10), LMU (3), UGOE/SUB (30), UdS (13), UHH (10)

Table 5.15 Deliverables and milestones for TA1 M4

#	Month	Deliverable
D1.4.1	12, 36, 60	Documentation of different procedures and tools
D1.4.2	24	Report on survey about community specific relevance of structural information in library and archive holdings
M1.4.3	30	Initial releases of annotation services
D1.4.4	48	Integration of annotations into annotation store
D1.4.5	54	Exemplary integration of components making explicit structural information relevant to the community (see D1.4.2)

TA1 M5: Community activities

Lead: UHH

Contributors: DIE, DNB, IDS, LMU, MWS, UGOE/SUB, UdS, UHH, UniTR, UzK

TA1 contributes to M5 as follows (see strategic theme 3, objective 6).

Activity 1 – Dissemination & outreach activities: TA1 will disseminate results from all measures at relevant conferences and events, such as the linguistic and DH conferences (e.g. DGfS, DHd) and annual meetings of learned societies of a range of disciplines. TA1 continues to network and coordinate with relevant specialized information services (SIS) (e.g. Romanistik, Linguistik, Philosophie, etc.), other NFDI consortia, CLARIN, and DARIAH

as well as the library and archive communities – nationally and internationally – and contributes TA-specific input to outreach activities (public relations). TA1 continuously evaluates its (web) services with respect to usability in new domains, e.g. by questionnaires and qualitative interviews. TA1 publishes experience-based reports, recommendations, and guidelines, especially regarding Measures M2-M4, to feed into best practices in research methods and research data management.

Activity 2 – Training & education: TA1 will make specific contributions to training events and educational materials in various formats (on-site and on-line workshops/training events, courses at summer schools, videos) in all TA-related areas described above, especially on LLM-based metadata annotation, LLM-supported transcription and annotation, technical (e.g. creating and using DTFs) and legal aspects such as readability, reconstructability, TDM exception in European copyright law, and linking data.

Activity 3 – Consulting, helpdesk & support: TA1 will contribute TA-specific support for groups and individuals (e.g. Text+ Helpdesk, Research Rendezvous) by expertise of the partners and participants, including consultation regarding legal and ethical issues, support in corpus development, annotation workflows, data curation, and support in infrastructure integration respecting the FAIR principles (e.g. Text+ Registry and – wherever possible – the FCS).

In-kind contributions by partners (in PM): DIE (6), DNB (24), MWS (20), UGOE/SUB (9), UdS (6), UHH (30)

Requested funding (in PM): IDS (4), LMU (9), UGOE/SUB (10), UdS (36), UHH (10), UniTR (12), UzK (6)

All achievements of TA1 in M5 correspond to the joint deliverables of M5 in section 5.0.

Cooperation with other task areas

The data domains (TAs 1-4, see sections 5.1-5.4) share common measures, objectives, and deliverables. The respective details are described in the introduction of the work programme (see section 5.0). Beyond the shared work in the joint deliverables, TA1 can be considered the “raw material” for any more structured form of research data. Hence, TA1 is the foundation for the other data domains, providing curated collections relevant for specific research questions involving corpora, editions, and dictionaries. Close cooperation with TA5 and TA6 is crucial for the success of TA1. Hence, TA1 centers will provide high quality metadata to describe their resources in the Text+ Registry as well as endpoints to the FCS with resources that have been enriched, e.g., with entity links and structural information. Furthermore, TA1 will contribute expertise in the Helpdesk – especially in legal and ethical issues – and will also actively support the work of the Communication Office and provide up to date and high-quality input to outreach activities and the Text+ web portal.

Table 5.16 Risks of implementation TA1

Risk identified	Risk prevention and mitigation
Insufficient scalability of services	Use cloud-based services and modular components that can be easily expanded. Monitor service performance and adjust dynamically.
Delay in certification processes	Install an internal certification working group. Offer support in Nestor (by DNB) and CoreTrustSeal (by CLARIN and DARIAH partners in TA1) certification and closely monitor progress.
Rapid evolution of AI-based tools	Maintain flexibility in workflows, monitor emergence of new methods and tools.
Quality limits of AI-based tools	Revert to/combine with other machine learning methods.
Lack of user uptake of services	Continuous interaction with communities (e.g. by SCC), collection of user feedback through surveys, workshops and community activities. Provide training and education materials.
Conflicting user demands	Monitor user demands and negotiate compromises.
Extensive requests in legal and ethical issues	Manage expectations, monitor dynamics of laws and regulations. Issue general guidelines and keep them up to date.

5.2 TA2: Language Corpora and Elicited Data

Co-leads: IDS, EKUT

Contributors: AdWHH, BAdW, BBAW, EKUT, FU, HAdW, IDS, LMU, RUB, UGOE/SUB, UniBI, UFR, UzK

Task Area 2 (TA2) focuses on collecting and curating diverse language corpora and elicited data, which are key resources in linguistics and literary studies as well as related fields where language serves as a primary medium of inquiry. Language corpora captures diverse aspects of human communication, while elicited data from experiments and observations allow the study of subtle, rare, and context-dependent phenomena. By integrating both data types, TA2 supports a broad spectrum of data, from everyday conversation and literary texts to experimental and neurocognitive studies.

TA2 contributes to Text+ RDM goals by operationalising several of its key pillars:

1. FAIR and CARE Principles guide all activities.
2. Data space integration ensures long-term preservation of language corpora and elicited data via certified Text+ Centers using tools like the Text+ Registry and FCS for discoverability and reuse.
3. Standardization and interoperability through contributions to (metadata) standards, aligning with the RDM strategy (see section 4.2), supporting machine-readability and processing across diverse data types.
4. Community-centered infrastructure through workshops, consulting, and cooperation projects, aligning infrastructure with scholarly needs.
5. Sustainable quality assurance via expert review, community validation, and regular updates to reflect evolving formats, tools, and research interests.

TA2 pursues the following three key objectives:

1. Ensure high data quality and reusability

To ensure high quality and reusability, TA2 focuses on: (1) machine-readable metadata and annotations; (2) addressing domain-specific challenges, like overlapping speech, consent in (CMC) data, or TEI variation; (3) promoting subfield-specific best practice guidelines; and (4) ensuring legal and ethical compliance. These efforts are embedded in the Text+ infrastructure, making resources discoverable via the Text+ Registry and the FCS.

2. Support for multimodal models

TA2 supports the integration and reuse of multimodal data – spoken language, facial expressions, physiological signals such as EEG or laryngography. It promotes models that align linguistic, visual, acoustic, and physiological data – enabling analysis of performance-based literature and audiovisual archives. TA2 advances multimodal analysis across disciplines and provides tools for seamless integration into the Text+ infrastructure.

3. Advance infrastructure for elicited data and experimental research

TA2 supports the collection, curation, and reuse of elicited language data essential for (psycho-)linguistic and literary research. It provides infrastructure for storing experimental data – such as surveys, eye-tracking, and neurophysiological measurements – from methods such as reading time studies and MRI. This TA ensures robust metadata and repository support for non-textual data, enabling discovery via the Text+ Registry. Hence TA2 supports multimethod research within a unified infrastructure.

The common objectives of the measures linking the four data domains of Text+ are described in detail in the introduction to the work programme (see section 5.0), also outlining the relevance of the respective measure to the NFDI.

TA2 M1: Federated research data infrastructure

Co-leads: EKUT, IDS

Contributors: AdWHH, BAdW, BBAW, EKUT, HAdW, IDS, LMU, UGOE/SUB, UzK

This measure ensures the sustainable operation and enhancement of Text+ Centers that manage diverse language corpora and elicited data. It focuses on infrastructure for spoken, multimodal, annotated, elicited, and historically structured corpora, with legal and ethical compliance as a core concern.

Activity 1 – Center-based infrastructure for diverse research data types: Maintains and develops infrastructure for: (1) spoken and multimodal corpora (e.g., endangered languages, sign languages); (2) highly annotated written language data (e.g. literary and linguistic corpora); and (3) Elicited Data.

Activity 2 – Metadata standardization and persistent identifiers: Implements standardized, machine-readable metadata (e.g. CMDI, ISO 24622-1 and ISO 24622-2) and assigns PIDs (ISO 24619), including entity linking (e.g. GND).

Activity 3 – Integration into the Text+ Data Space: Connects all corpora and elicited data resources via: (1) FCS-compatible endpoints; (2) registry integration via automated metadata sharing; and (3) maintenance of the registration with re3data.org.

Activity 4 – Quality assurance and certification ensuring that participating centers: (1) maintain or pursue certification (e.g., CoreTrustSeal, nestor); (2) implement quality control workflows for ingestion and curation; and (3) report on data sets (PID + metadata + data).

Activity 5 – Legal and ethical compliance addressing the legal and ethical challenges, e.g.: (1) respect of community rights for indigenous and minority language data; (2) compliance with GDPR and national data protection rules, including finding an appropriate legal basis for processing; (3) compliance with copyright legislation.

Activity 6 – Third-party data integration and depositing services: This activity supports the ingestion of external data sets.

Relation to TA2 objective 1 (Data Quality and Reusability), objective 2 (Support for Multimodal Models), and objective 3 (Elicited Data Infrastructure).

RDM aspects: Long-term preservation, FAIR compliance, repository certification, legal/ethical handling, metadata harmonization, interoperability, PID assignment, federated access, data curation, and community engagement.

Deliverables for TA2 M1

Common deliverables for M1 of the data domains are: D0.1.1(see section 5.0).

All results and achievements of M1 in TA2 correspond directly to the joint deliverables of M1 as defined in 5.0.

In-kind contribution by partners (in PM): AdWHH (30), BAdW (18), BBAW (30), EKUT (30), UGOE/SUB (15), HAdW (9), IDS (30), LMU (60), RUB, UzK (22)

Requested funding (in PM): -

TA2 M2: Interfaces between data from data domains

Co-leads: EKUT, IDS

Contributors: AdWHH, BBAW, EKUT, IDS, LMU, UGOE/SUB, UzK

This measure develops interfaces to enable semantic interoperability between language corpora, elicited data, and other Text+ domains. It focuses on integrating metadata and object data through a scalable, machine-actionable semantic layer for cross-domain retrieval and analysis.

Activity 1 – Knowledge graph representation for metadata and object data: Implements Linked Data standards (e.g., RDF, OWL (W3C, n.d.a), and the Web Annotation

Data Mode) to represent metadata and explore object data integration. It includes: (1) SPARQL endpoints for federated querying; (2) integration of external knowledge sources such as GND, Wikidata, DBpedia (DBpedia Association, 2025), and NFDI4Culture's Culture Knowledge Graph; and (3) PID resolution for stable linking.

Activity 2 – Semantic enrichment and authority file integration: Integrates automated NLP pipelines (e.g., MONAPipe, WebLicht (Hinrichs et al., 2010) and LLM-based methods to enrich object data (e.g., TEI-XML) and metadata, e.g. for tasks such as named entity recognition, disambiguation, and alignment with authority files (e.g., GND, Wikidata). It focuses on: (1) applying and extending the Web Annotation Protocol Compliant Annotation Store (WAPCAS) with public APIs for annotation delivery (e.g., to FCS), support for annotations across text, audio, video, and other binary formats; and (2) support for ontology integration.

Relation to TA2 objective 1 (data quality and reusability), objective 2 (support for multimodal models), and objective 3 (elicited data infrastructure).

RDM Aspects: Long-term preservation, FAIR/CARE compliance, infrastructure interoperability, data curation, quality assurance, repository services.

In-kind contribution by partners (in PM): EKUT (24), IDS (30), UGOE/SUB (3)

Requested funding (in PM): AdWHH (24), BBAW (6), EKUT (36), IDS (38), LMU (18), UGOE/SUB (21), UzK (15)

Table 5.17 Deliverables and milestones for TA2 M2

#	Month	Deliverable/Milestone
D2.2.1	24	Implementation of SPARQL endpoints and persistent identifier resolution
D2.2.2	36	Integration of NLP pipelines and LLM-based enrichment tools
D2.2.3	48	Deployment of WAPCAS APIs and annotation service
D2.2.4	60	Reports on semantic integration use cases and cross-domain query capabilities
M2.2.5	48	Semantic integration by NLP workflows for metadata and object data production ready

TA2 M3: Enabling text and language data for AI

Co-leads: EKUT, IDS

Contributors: AdWHH, BBAW, EKUT, IDS, LMU, UGOE/SUB, UzK

This measure explores the two-way relationship of language corpora and elicited data with AI technologies, particularly large language models (LLMs). It aims to enable responsible AI development using corpora and elicited data, and integrate AI methods into data annotation, enrichment, and retrieval, with a focus on transparency, reproducibility, and ethics.

Activity 1 – Preparing corpora for AI training and fine-tuning: This activity converts existing corpora into formats suitable for training and fine-tuning domain-specific LLMs. It

includes: (1) identifying and preparing corpora containing non-standard, historical, or dialectal language; (2) supporting the creation of specialized transformer or embedding models; and (3) reducing bias and improving inclusivity in AI outputs.

Activity 2 – Benchmarking and evaluation for humanities-specific AI tasks: This activity develops benchmarks and protocols for humanities-focused AI-tasks, promoting: (1) transparent and replicable workflows; (2) open evaluation designs that respect data protection and copyright; and (3) responsible AI practices in academic contexts.

Activity 3 – Repository support for AI-generated resources: This activity ensures long-term preservation of AI-generated or fine-tuned models, with: (1) metadata profiles covering authorship, provenance, and AI-specific attributes; and (2) long-term archiving of models and associated data, including derived text formats.

Activity 4 – AI-assisted annotation and enrichment workflows: This activity integrates LLMs into annotation pipelines to support tasks such as: (1) glossing, semantic role labeling, syntactic analysis; (2) hybrid workflows combining LLMs with tools like MONAPipe or WebLICHT; and (3) controlled vocabulary extraction and LOD linking.

Activity 5 – Retrieval-Augmented Generation (RAG) and semantic search: This activity develops services for semantic search, including key aspects summarized for brevity.

Relation to TA2 objective 1 (Data Quality and Reusability), objective 2 (Support for Multimodal Models), and objective 3 (Elicited Data Infrastructure).

RDM Aspects: Long-term preservation, FAIR compliance, metadata harmonization, interoperability, federated access, quality assurance.

In-kind contribution by partners (in PM): EKUT (24), IDS (30), UGOE/SUB (3)

Requested funding (in PM): AdWHH (21), BBAW (6), EKUT (30), IDS (38), LMU (18), UGOE/SUB (18), UzK (15)

Table 5.18 Deliverables and milestones for TA2 M3

#	Month	Deliverable/Milestone
D2.3.1	24	Workflow to convert speech and text corpora in repositories to AI-ready formats
D2.3.2	36	Training materials and user documentation to support the generation of AI-compliant corpora
D2.3.3	48	Web services, tools, and APIs integrating AI-enabled data retrieval, transcription, and augmentation
D2.3.4	24	Extended metadata profiles covering critical AI issues (e.g. authorship, data provenance)
D2.3.5	24, 36, 48, 60	Annual research data management workshops on AI and language data
M2.3.6	36	Integration model for AI generated data

TA2 M4: Implicit and explicit structures of language

Co-leads: EKUT, IDS

Contributors: BBAW, EKUT, IDS, UGOE/SUB

This measure focuses on enriching language corpora and elicited data by making implicit linguistic structures explicit. Elicited data often includes prompts designed to elicit specific phenomena (e.g. syntax, phonetics, morphology) which require targeted annotation to be machine-readable. Similarly, textual and spoken corpora can be structurally enriched using NLP tools (see for example WebLICHT, SpaCY, MONAPipe) to reveal syntactic, discourse, and stylistic features. Time-aligned transcriptions in spoken corpora can be transformed into ISO 24624-compliant formats. Advanced tools like KorAP and TüNDRA enable deep querying of such enriched data beyond surface-level search.

Activity 1 – Reference implementations for multi-layered annotation: This activity develops and integrates tools such as KorAP for exploring and comparing multiple annotation layers supporting: (1) structural alignment and navigation; and (2) support for coexisting structural interpretations.

Activity 2 – User-centered documentation and training: This activity provides training materials and best-practice guides, including key aspects summarized for brevity.

Activity 3 – Semi-automated annotation using reasoning-based LLMs: This activity explores language models for inferring annotations in small or TEI-encoded corpora, combining: (1) linguistic reasoning with lexical resources and analogous corpora; and (2) expert-guided, context-aware annotation.

Activity 4 – Tools and APIs for structural enrichment: This activity develops tools and APIs for managing annotation layers, for: (1) structural enrichment of corpora; (2) integration with existing annotation frameworks; and (3) interoperability with other Text+ services.

Relation to TA2 objective 1 (Data Quality and Reusability), objective 2 (Support for Multimodal Models), and objective 3 (Elicited Data Infrastructure).

RDM Aspects: Long-term preservation, FAIR compliance, interoperability, federated access, searchability, infrastructure interoperability, data curation, quality assurance.

In-kind contribution by partners (in PM): EKUT (12), IDS (18), UGOE/SUB (3)

Requested funding (in PM): BBAW (6), EKUT (12), IDS (15), LMU (6), UGOE/SUB (18), UzK (6)

Table 5.19 Deliverables and milestones for TA2 M4

#	Month	Deliverable/Milestone
D2.4.1	12	Reference implementations using KorAP or other tools for multi-layered corpus exploration
D2.4.2	24	User documentation and training materials for identifying and annotating implicit structures
D2.4.3	36	Tools and APIs for adding and aligning annotation layers in corpora
D2.4.4	48	Annotated corpora demonstrating transitions from implicit to explicit structure

D2.4.5	60	Reports on structural diversity and annotation practices in language corpora
M2.4.6	48	Support for multiple structures production ready

TA2 M5: Community activities

Co-leads: EKUT, IDS

Contributors: AdWHH, BAdW, BBAW, EKUT, HAdW, IDS, LMU, UniBI, RUB, UGOE/SUB, UzK

This measure provides consulting, training, and outreach for researchers working with language corpora and elicited data. It promotes FAIR- and AI-readiness (FAIR-R), responsible AI use, and sustainable RDM practices, while enhancing the visibility and uptake of Text+ services.

Activity 1 – Consulting and data depositing services: This activity offers–integrated into the Text+ Helpdesk–expert consultation on: (1) corpus building, data curation, and data management planning; (2) ethical and legal aspects; (3) FAIR- and AI-readiness; and (4) depositing workflows.

Activity 2 – LLM-based support assistants: This activity develops interactive assistants to: (1) replace static FAQs with dynamic, context-aware guidance; and (2) support RDM and FAIR-R.

Activity 3 – Training events and educational formats: This activity organises regular training events for people who use this type of data.

Activity 4 – Training material development and publication: This activity produces Open Educational Resources (OER) in this data domain.

Activity 5 – Dissemination and community engagement: This activity promotes TA2 services and outcomes via (1) conferences, workshops, and scholarly networks; (2) coordination with FIDs, and Data Competence Centers , and other Text+ task areas.

Relation to TA2 objective 1 (Data Quality and Reusability), objective 2 (Support for Multimodal Models), and objective 3 (Elicited Data Infrastructure)

RDM Aspects: Community involvement, consulting, training, helpdesk, RDM

In-kind contribution by partners (in PM): EKUT (12), IDS (12)

Requested funding (in PM): BBAW (12), EKUT (12), IDS (18,63), LMU (9), UGOE/SUB (3), UzK (9)

All results and achievements of M5 in TA2 correspond to the joint deliverables of M5 in 5.0.

Cooperation with other task areas

TA2 collaborates closely with other data domains and TA5 (Infrastructure/Operations) through the shared implementation of measures M1–M5. In addition to this structural

alignment, TA2 contributes domain-specific expertise to foster integration and innovation across Text+.

Cross-task area co-operation (beyond common measures):

- With TA1 (Library Holdings as Research Data): Joint work on integrating annotated literary corpora from digitized library data, including metadata harmonization and NLP-based enrichment.
- With TA3 (Lexical Resources): TA2 provides corpus-derived data (e.g., frequency, collocations) to enrich lexical databases, which in turn support annotation and disambiguation.
- With TA4 (Editions): Collaboration on linguistic annotation in digital editions and the use of LLMs for structural and semantic enrichment.

The cooperation with TA5 is essential to TA2's success. Key areas of collaboration include: (1) FCS: TA2 contributes endpoints and supports integration of multimodal and elicited data; (2) Metadata and Registry services: Joint work on integrating TA2 metadata into the Text+ Registry; (3) LLM services and RAGestry: Collaboration on AI-enhanced interfaces for language data access; and (4) Helpdesk and Consulting: TA2 provides domain-specific support on FAIR/AI readiness, data depositing, and legal/ethical issues.

Table 5.20 Risks of implementation TA2

Risk identified	Risk prevention and mitigation
Heterogeneity of data types and annotation standards across corpora and elicited data	Develop and promote guidelines for various data types; implement metadata harmonisation strategies also drawing back on NFDI-wide discussions.
Legal and ethical complexities in handling data affecting third party rights (e.g., spoken, CMC, elicited, corpora)	Provide consultation services; integrate legal review into data workflows; ensure GDPR and copyright compliance.
Limited availability of training data for AI applications in low-resource domains	Use hybrid approaches combining LLMs with curated resources; support semi-automated annotation and enrichment; create training material and tutorials, including creating material assisted by AI methods (e.g. by using Synthesia (Synthesia Ltd., n.d.) or similar generators).
Technical bottlenecks in infrastructure integration, particularly in registry and FCS connectivity with data sets provided in TA2 M1	Promote automation and scalability in registry workflows; establish clear responsibilities and transparent reporting mechanisms across centers in collaboration with TA Administration; provide technical support and shared documentation to reduce manual workload.
Rapid evolution of AI tools and standards	Maintain flexibility in workflows; evaluate and integrate emerging tools; contribute to responsible AI guidelines.

5.3 TA3: Lexical Resources

Co-leads: BBAW, UMR

Lexical resources encompass reference works and databases focused on word usage. Their design and compilation are a research endeavour yielding significant contributions to

fields such as lexicography and semantics. The resulting resources serve both research and documentation needs across various domains. Key types of lexical resources include a) dictionaries for human readers, covering diverse languages, regional varieties, historical periods, and specialized domains; b) machine-readable dictionaries; c) printed or digitally born dictionaries and encyclopaedias; d) terminological databases tailored to specific knowledge areas; e) ontologies that organise lexical content according to conceptual systems or semantic relationships; f) wordlists, including frequency lists or linguistically annotated lists (morphology, syntax, semantics); g) word maps and linguistic atlases.

From an IT and infrastructural perspective, the domain of lexical resources includes several key elements: infrastructure units like platforms and dictionary portals for accessing, searching, and storing lexical data, software services for processing this data, and established best practices and standards for metadata, object data, and data interchange.

In Text+, the primary user groups for lexical resources are researchers and research projects in all scientific disciplines involved in lexical research and documentation, technical systems, including lexical information systems and (large) language models. These groups each require specific services and support that are tailored to every stage of the data lifecycle, including production, usage, and research data management.

In the first phase of Text+, the TA Lexical Resources has successfully addressed major challenges. These included the unification of resources within a diverse landscape of lexical resources and assisting data providers in the FAIRification and CAREification processes (refer to section 2.1, Grand Challenge I and II). This was accomplished by implementing a shared metadata scheme and harmonizing it with the schemas of all TAs. A common and interoperable format standard (TEI Lex-0) and a query system (Lexical Federated Content Search (LexFCS), Körner et al., 2023) have been established as part of the broader Text+ FCS. A user interface (LexDataView) harmonizes access to lexical data across various resources provided by the data centers participating in the TA.

The second project phase will build upon the successes of the first phase. All existing infrastructure components will be maintained and further developed (M1). Concurrently, new methods for harmonizing, mapping, and querying lexical data will be investigated and implemented, aligning with the strategic themes outlined in section 2. Resource harmonization will expand to include new relation types, such as geographical relations, and resource mapping will increasingly adopt a cross-domain focus (M2). A key focus will be set on making specialized knowledge within lexical resources accessible for machine learning approaches like large language models (M3), directly addressing the first strategic theme (“language/text as dominant interface”). The second strategic theme (“dynamic interplay of structured/unstructured data”) is primarily addressed by M4, which will explore the dynamic interaction between structured and unstructured data and develop

infrastructure for structuring lexical resources. M5, with its overarching scope, enables a comprehensive approach to the third strategic theme ("human-centered research culture") across all other measures and the entire lexical resource domain: fostering and maintaining broad research community connections will ensure that the research culture promoted by the TA Lexical Resources remains human-centered and responsive to the needs of researchers in the field.

For its second phase, Text+ will integrate 10 major stakeholders operating German data and competence centers that offer lexical, lexical-semantic, ontological, related resources, and training and consulting. Eight of these centers were part of Text+ during its initial phase: **BBAW (Co-Lead)**, IDS, UniTR, EKUT, SAW, UzK, BAdW, **UMR (Co-Lead)** (since 2023). The following two centers will join the TA for the second phase as participants: UniRo, AdWHH.

The selection of these data centers was based on several criteria: sufficient technical maturity, hosting of major lexical resources, the ability to address representative user groups across various communities, a compelling proposal for added value, and sustainable in-kind contributions to support the offering.

BBAW and UMR will jointly lead the TA Lexical Resources. BBAW will focus on the TA-specific work plan and internal cooperation across data domains (M1–M4). UMR will primarily manage external collaborations with the research community (M5).

The common objectives of the measures linking the four data domains of Text+ are described in detail in the introduction to the work programme (see section 5.0).

TA3 M1: Federated research data infrastructure

Co-Lead: EKUT

Contributors: BBAW, SAW, UMR, IDS, UniTR, BAdW, AdWHH, UzK

Throughout the first project phase, all partners have operated their data centers for lexical research data within the Text+ infrastructure. This commitment will persist into the second phase, with individual centers maintaining responsibility and using their matching funds for the general infrastructure.

Activity 1: M1 facilitates technical harmonization, collaboration, and joint development of the Text+ infrastructure. Key areas of focus include data center certification, data archiving, metadata provision, data ingest for lexical resources, and maintenance of LexFCS interfaces for partners' resources. This inter-domain collaboration, in conjunction with TA IO, is vital for advancing the Text+ infrastructure and establishing the Text+ data space. Consequently, all partners managing data centers will participate in cross-domain collaboration to ensure their software implementations align with Text+ data space specifications.

Activity 2: Partners in the TA will continue to actively engage in international standardization bodies such as ISO (EKUT, IDS), TEI (BBAW, UniTR, IDS), and TEI Lex-0 (BBAW) to ensure close alignment with current developments in infrastructure specifications and in data formats for lexical resources.

Activity 3: Some data centers will use matching funds that go beyond general infrastructure maintenance and development. Namely, BBAW and SAW will also ensure the ongoing provision of LexFCS endpoints for cooperation projects completed in phase I. Additionally, AdWHH (Gebärdensprache), BAdW (dialect portal), and UzK (Kosh) will implement new LexFCS endpoints for their respective lexical data.

In-kind contribution by partners (in PM): BBAW (45), SAW (18), EKUT (36), UMR (30), IDS (30), UniTR (15), UzK (32), BAdW (12), AdWHH (8)

Requested funding: -

Deliverables and milestones: All results and achievements of M1 in TA3 correspond directly to the joint deliverables of M1 as defined in section 5.0.

TA3 M2: Interfaces between data from data domains

Co-Lead: BBAW

Contributors: SAW, EKUT, UMR, IDS, UniTR

M2 aims to establish cross-domain interfaces from lexical resources to various *external* data. This builds upon the achievements of the first project phase, which concentrated on interfaces *among* lexical resources provided by the partners of the TA. A key prerequisite for this endeavour is the implementation of a scheme for persistently identifying and addressing distinct lexical data, particularly individual entries and meanings.

By the end of the first project phase, the LexFCS is able to provide researchers with means of harmonized querying of lexical resources. In the second project phase, Text+ will significantly extend the ways to query its lexical resources, most notably through knowledge graph embeddings and by implementing direct access for state-of-the-art artificial intelligence tools such as large language models (M3). The use of persistent identifiers will enable the other data domains to exploit lexical resources for common tasks such as semantic annotation. Within M2, two main goals can be identified: linking of entities across resources (activity 1) and providing public linking services to the community (activity 2).

Activity 1: Persistent identification of (lexical) information within a resource allows for sustainable references across all data domains. All partners will introduce a mechanism for persistent identification for their resources, on the level of individual entries and, if available, on the level of word meanings. The LexFCS will be extended for use with persistent identifiers on entry (denotation) level. (M3.2.10)

Building on initial experiments of resource linking in the context of (linguistic) linked open data ((L)LOD) by SAW and EKUT, and on a close collaboration between the DWDS and Wikidata which the BBAW will provide in-kind, cross-domain linking on a broader and domain-independent level will be pursued in M2. A report on the evaluation of knowledge graph representations for lexical resources will be provided (D3.2.2). Based on this report, selected lexical resources will be transformed into LOD compliant formats, e.g. via a SPARQL endpoint, including retrodigitized and born-digital resources as well as geolinguistic and multilingual data (M3.2.4). An alignment between the DWDS and GermaNet will be created in this framework effectively creating a cross-resource knowledge graph. A query interface for lexical knowledge graphs will be created that provides another, more generic interface to the lexical data in the domain besides the LexFCS (M3.2.11). Providing our lexical resources via LOD aware endpoints also enables queries and linkages from data domains outside the Text+ consortium, effectively creating a generic interface, e.g. for other consortia within the NFDI. To foster the inclusion of geolinguistic data, a unified reference system for geolinguistic metadata will be created (M3.2.8, M3.2.9).

Activity 2: A service for interoperable and context-aware cross-domain word sense disambiguation, annotation, and entity linking will be designed and implemented in M3. This service critically depends on comprehensive and persistently identifiable lexical resources (as described in 1 above). Given a key-word and a context (e.g. a sentence), this service will respond with all entries and/or senses in lexical resources that exhibit the correct meaning of the key-word. Part of the disambiguation service will be a lemmatiser of modern and historical (16th–19th century) German forms. Based on their REDE-SprachGIS (Schmidt et al., n.d.) and other sources such as GND files, resources of the “Deutsche Biographie” or of the “Digitales Familiennamenwörterbuch Deutschlands”, a curated lists of named entities (local names, person names) will be created that may serve as valuable input for the entity linker proposed in M3. Moreover, standardized historical and current place and spatial data will be delivered, thus enabling linking to related information within the Text+ infrastructure. This includes unified geospatial metadata, data formats (e.g. *.geojson), geometries (e.g. polygons, point layer), and connections to linguistic and regional data.

Text+ will design and implement a service to generate lexico-statistical resources (e.g. word lists with word frequencies) from corpus resources that do not allow free reuse. These lists are useful in education, clinical research, and psycholinguistics (e.g. case study wordlists for text classification, D3.2.5).

In-kind contribution by partners (in PM): BBAW (30), SAW (6), UMR (45), IDS (10)

Requested funding (in PM): BBAW (30), SAW (20), EKUT (25), UMR (15), IDS (15), UniTR (8)

Table 5.21 Deliverables and milestones

#	Month	Title
D3.2.1	12	Concept for persistent identification within lexical resources
D3.2.2	12	Overview of knowledge graph representations for lexical resources
D3.2.3	18	Concept for linking geolinguistic ressources
M3.2.4	36	Provision of lexical resources as knowledge graphs
M3.2.5	36	Tool for the generation of lexico-statistical resources
M3.2.6	48	Persistent identification of lexical resources on entry/sense level
D3.2.7	48	Evaluation of methods and models used for creating knowledge graphs
M3.2.8	48	Provision of standardized reference system for geolinguistic metadata of lexical resources
M3.2.9	60	Provision of geometries for geo-entities within lexical resources
M3.2.10	36	LexFCS extended to work with persistent identifiers
M3.2.11	48	Query interface for lexical knowledge graphs
M3.2.12	48	Service for cross-domain word sense disambiguation, annotation, and entity linking

TA3 M3: Enabling text and language data for AI

Co-Lead: UMR

Contributors: BBAW, SAW, EKUT, IDS

Lexical resources, including dictionaries, thesauri and ontologies, as well as various other structured linguistic databases, represent a highly sophisticated form of data. Despite their inherent value and the meticulous effort invested in their creation, these resources have, until now, been largely underutilized in the training of LLMs. This oversight has significant implications, particularly for under-resourced languages or language varieties. This scarcity of training data leads to several critical issues, including limited performance, bias, and inaccuracy, and in the long term it may lead to an erosion of linguistic diversity.

One way to overcome these challenges is the integration of comprehensive lexical resources into LLM training. This integration enhances LLM understanding across various languages by providing structured linguistic information on vocabulary, grammar, semantics, and pragmatics. The benefits include improved LLM performance for under-resourced languages, bias mitigation through representative data sets, and inaccuracy and increased linguistic inclusivity in the AI landscape.

Within M3, two main goals can be identified: the transformation of lexical data to foster its use in AI applications (data readiness, activity 1), and semantic annotation services that will rely on AI procedures (e.g. entity and sense linking, activity 2).

Activity 1: For large language models (LLMs) to effectively leverage the rich information contained within lexical resources, a critical prerequisite is their accessibility through standardized and widely adopted interfaces such as the Model Context Protocol (MCP). Currently, lexical resources are provided in disparate formats or locked within proprietary

systems, hindering seamless integration with advanced AI applications. MCP allows LLMs to directly query, parse, and incorporate lexical information, such as word senses, syntactic, or valency structures, semantic relationships, and morphological variations, into their internal representations. The lack of such unified access points currently necessitates complex and often lossy conversion processes or forces LLMs to rely on less precise inferential methods. Therefore, the integration of lexical resources with established protocols like MCP is paramount for fostering more nuanced, accurate, and contextually aware LLM capabilities, ultimately leading to significant advancements in natural language processing applications.

A service for the spatial mapping and modelling of dialectal language variants will be developed, which enables the geolocation of written texts and speech samples. This service facilitates the preparation and georeferencing of dialect data, allowing regional linguistic features to be assigned to specific geographic areas. The resulting geolocated language resources are stored in a continuously growing database and provide high-quality training data for AI applications with a geolinguistic focus.

Activity 2: A semantic annotation service that links word occurrences in full text resources to sense definitions will be designed and implemented (M3.3.5, M3.3.8). Using transformer-based approaches, context information will be used for sense disambiguation, and thus augmentation and enrichment via external knowledge sources like GermaNet (Zinn et al., 2022), DWDS (Klein & Geyken, 2010), and other lexical resources will be provided.

A translation assistant for dialect texts (translation memory) will be contributed that supports the interpretation and partial translation of regional and dialectal vocabulary through semi-automatic suggestion logic. It will be based on the database for the spatial mapping service. The system generates lexical parallel data to improve the machine readability of dialect texts, support future AI-based translation technologies, and promote interoperability with digital editions and language processing services (M3.3.10).

In-kind contribution by partners (in PM): UMR (45), IDS (15)

Requested funding (in PM): BBAW (30), SAW (20), EKUT (25), UMR (20), IDS (20)

Table 5.22 Deliverables and milestones

#	Month	Title
D3.3.1	12	Concept for a semantic annotation service
D3.3.2	12	Concept and prototypical implementation of MCP-Server/Client
D3.3.3	12	Concept for structured and georeferenced dialectal language data as AI training resources
D3.3.4	12	Concept for spatial embedding and interpolation of dialectal variation using georeferenced data
M3.3.5	36	Prototype of a semantic annotation service
M3.3.6	36	Provision of georeferenced, structured dialectal data for AI training purposes
D3.3.7	48	Report/Publication on technical and conceptual implications of KI/MCP integration of lexical resources
M3.3.8	54	Final version of a semantic annotation service
D3.3.9	60	Guidelines/best-practices for MCP resource integration process
M3.3.10	60	Deployment of a translation assistant enabling AI applications in dialect-to-standard lexical mapping

TA3 M4: Implicit and explicit structures of language

Co-Lead: EKUT

Contributors: BBAW, SAW, UniTR, UzK

Lexical resources are highly structured models of languages designed to condense lexical and linguistic information as much as possible, while remaining representative and consistent. A key objective in their design is to present linguistic information in explicit structures. However, many dictionaries and encyclopedias, originally printed and not yet digitized, lack a layer of explicit lexical information.

Activity 1: As an immediate implementation of the second strategic theme (see 2. ST2)–the dynamic interplay between structured and unstructured data–Text+ will provide the community with a comprehensive documentation of currently available tools (e.g. elexifier) and established procedures for the generation of structured lexical resources (D3.4.1).

To encourage using and exploring currently unstructured lexical resources, such as those obtained through OCR digitization, a conversion service will be implemented (M3.4.2). This service will provide researchers with the necessary infrastructure to curate raw text data into a Text+ compatible format which can then be integrated into the LexFCS. Initial efforts will focus on leveraging existing libraries of conversion routines previously employed by partners in various conversion projects. Additionally, a classification system for levels of lexical structuring and annotation will be developed. This classification will function as a metadata entity, enabling the location and retrieval of resources based on their specific explicit structures and structuring depths.

Transforming the structure of a resource may entail changes in the explicitness of annotations or even the structure itself. Consequently, an evaluation of the methods and

models used for creating knowledge graphs as proposed for M2 will be needed to assess the impact of such transformations on the structure of lexical resources (D3.4.3).

In-kind contribution by partners: -

Requested funding (in PM): BBAW (5), SAW (10), EKUT (25), UniTR (5), UzK (18)

Table 5.23 Deliverables and milestones

#	Month	Title
D3.4.1	24	Overview of available tools and established procedures for the generation of structured lexical resources
M3.4.2	48	Tools and workflow for (semi-)automatic structuring of unstructured full text
D3.4.3	60	Evaluation of methods and models used for creating knowledge graphs

TA3 M5: Community activities

Co-leads: UMR

Contributors: BBAW, SAW, EKUT, UniRo, IDS, UniTR, UzK

Close collaboration with the community is of particular importance to the TA, as the various disciplines place specific demands on the infrastructure regarding lexical resources within Text+. To best engage with the community and serve its needs, a dedicated Data Steward position will be established, supported by all partners. The tasks within M5 relate to three areas.

Activity 1 – Dissemination and outreach activities: Partners in the TA will participate in Text+ dissemination activities and actively promote Text+ services–especially lexical services as e.g. the LexFCS–at the annual conferences of relevant scholarly associations. Specific technical standards, services, and tools will be disseminated in European research infrastructures such as CLARIN and DARIAH by the SAW.

Proactive engagement with new communities will be a key focus of the second phase. Many projects of different sizes are already clear candidates for contact with the TA because their data structures require a conceptual advancement of the Text+ infrastructure. They provide ideal means to further develop our tools and services while guaranteeing that the necessary adaptations are of immediate impact for our community. Examples include resources containing geospatial information (e.g., “Digitales Familiennamenwörterbuch Deutschlands”) or associated image data (e.g. “Runische Schriftlichkeit in den germanischen Sprachen”), terminological resources, historical reference works, and encyclopaedias (e.g., “Deutsches Rechtswörterbuch”, “DTA”, “Deutsche Biographie”). In addition, born-digital (e.g. “FrameNet-Konstruktikon”) and graph-based (e.g. “ALMA – Wissensnetze in der mittelalterlichen Romania”) resources are also considered. Alongside European resources that implement a range of lemmatization strategies (e.g. “Verba Alpina”, “Frühneuhochdeutsches Wörterbuch”, "Wortbildung diachron", and the lexical

resources of the “Sorbisches Institut”), resources in non-Latin scripts are also considered (e.g. “Zoroastrian Middle Persian Digital Corpus and Dictionary”).

Beyond that, the UzK will reach out to typologists and field linguists collecting lexical data, often as a side product, to address the needs of these researchers, host their data, and integrate these resources into the federated research data infrastructure. The TAs focus is not restricted to large and well-known resources. Smaller data sets, such as highly specialized terminology lists or lexicographically innovative prototype dictionaries, are also of interest as they may challenge the data modelling and broaden the infrastructure’s capabilities to accurately represent them.

Activity 2 – Training and education: As in phase I, community workshops on the tools and infrastructure provided within the TA will be a primary forum for contact with our community. Text+ will engage in the collection and provision of resources for the research community regarding data and metadata curation and their structuring (see M4), in collaboration and alignment with the SSHOC Marketplace and EMLex. Fields of collections will be training materials, publications, and documentations of digitization and annotation workflows, documentation of tools and services, and the provision of code where applicable.

Activity 3 – Consulting, helpdesk and support: A data steward serves as a permanent access point for community interaction. They will support and actively participate in the Text+ helpdesk in matters concerning lexical resources, services, and workflows, thus providing first level support for the TA. A process-oriented approach will be adopted for consulting cases involving external contributors, providing a single point of access while involving domain specialists where necessary.

In-kind contribution by partners (in PM): UMR (15), UniRo (12), IDS (5)

Requested funding (in PM): BBAW (10), UMR (70), SAW (10), EKUT (15), IDS (7), UniTR (2), UzK (18)

> All achievements of M5 in TA3 correspond to the joint deliverables of M5 in section 5.0.

Cooperation with other task areas

The data domains and therefore the respective TAs (see sections 5.1–5.4) share common measures with common objectives and deliverables. The respective details are described in the introduction of the Work Programme (see section 5.0).

The development of the LexFCS will be closely aligned and harmonized with TA 5 (IO) to ensure seamless integration in the Text+ Webportal. Further development of the metadata model will be aligned with all data domains and TA 5 to ensure compatibility with the Text+ Registry. TA1 (LaRD) will serve as a possible provider of plain text of dictionaries and

encyclopaedias for further curation in M1 and M4. TA 3 will align its activities with the communications office (TA 6, M4).

Table 5.24 Risks of implementation TA3

Risk identified	Risk prevention and mitigation
Incompatible development of future LexFCS – old endpoints not integrated in up to date infrastructure	Provide help and support for migration of external endpoints.
(External) partners of federated infrastructure do not recognize latest developments and thus the need for upgrading, e.g. FCS endpoints	Provide subscription to mailing list regarding news of technical (breaking) changes.
Important external resources cannot be curated and integrated due to limited capacities in M1	Promote flexible fund applications to external stakeholders, provide help and support for setting up external data curation funding requests.
Rapid evolution of AI tools and protocols	Maintain flexibility in workflows, monitor emergence of new tools and protocols, align with partners' institutional AI guidelines.

5.4 TA4: Editions

Co-leads: AWK-NRW, TUDa

Within the spectrum of research data structuring, scholarly editions represent one of the most structured forms and models. As critical presentations of cultural heritage, they establish resources for scholarly reference, fulfilling a pivotal role in the humanities and beyond. Scholarly editions ensure the robust preservation, presentation, and contextualisation of sources, especially texts. Their level of annotation, commentary, and critical representation can adhere to different levels of detail and focus: from individual works to corpora, from hyper-diplomatic transcription to normalized renderings, from tracing the genesis of sources to collating their transmission variance.

This diversity of methods and materials is reflected in the field of research traditionally identified as textual scholarship. The computational turn towards digital scholarly editing adds further complexity: questions about best practices and standards, fundamental challenges of sustainability across all data layers, and, particularly in light of developments in AI, the impact of recent innovations and disruptive technologies. Discussions around the use of LLMs and ML-augmented tools in workflows and interface designs already dominate the field and will likely continue to do so in the (near) future.

TA4 addresses this dynamic situation with a participatory approach in order to facilitate the standardization of practices in the community. The key factor in this is the implementation of FAIR/CARE principles across a heterogeneous data domain with a focus on community activities. This also shapes the major contributions of TA4 to the Grand Challenges, objectives, and measures of the second funding phase: continuous integration and curation of edition metadata into the Text+ Data Space (M1), further dissemination and interoperability of the Text+ Registry as well as the Text+ FCS and their establishment as

reference systems (M2), guidelines and recommendations for LLM- and AI-driven workflows and sustainability challenges (M3), a framework for the technical implementation of complex data ecosystems and the transformation processes between explicit and implicit knowledge structures (M4), literacy campaigns and education efforts through the organisation of events, training materials, tutorials, and the provision of RDM support across the entire data-lifecycle of editions within the context of Text+ Consulting and Helpdesk (M5).

The common objectives of the measures linking the four data domains of Text+ are described in detail in the introduction to the work programme (see 5.0). Those descriptions, in addition, outline the relevance of the respective measure to the NFDI, indicators of success, as well as joint deliverables.

TA4 M1: Federated research data infrastructure

Co-Lead: AWK-NRW, HAB

Contributors: AWLM, BBAW, HAdW, h_da, KSW, MWS, NAWG, SAW, STI, UGOE/SUB, TUDa

All Text+ Centers in TA4 will continuously contribute to the further development, maintenance, and expansion of the federated research data infrastructure. With the aim of improving data quality within the framework of the FAIR criteria, an open registry for scholarly editions was successfully developed as part of the Text+ Registry in phase I (Gradl et al., 2024). This was achieved in close cooperation with TA5 on the basis of a dedicated data model (Text+ Task Area Editions, 2024), which allowed editions of TA participants to be integrated and also to aggregate catalogue metadata (e.g. (Sahle, 2020), (Franzini, 2022), (*AGATE – A European Gateway for the Academies of Sciences and Humanities*, n.d.)) to be curated and enriched. In addition, the Text+ Registry lists editing tools and research software included in relevant research software directories (nfdi.software, SSH Open Marketplace).

Activity 1 – Contribution and curation of metadata: In funding phase I, rich entries for all resources provided by participants were integrated into the Text+ Registry. Moving forward, all TA participants will continue to curate and update their data sets continuously. Ensuring the quality and up to dateness of entries via curation (including imported catalogue data) as well as a continued documentation are essential for the adoption of the Text+ Registry by the research community and its establishment as a tool of reference. The realization of these tasks is coordinated and carried out by an editorial team. Accompanying activities to increase acceptance and dissemination of this key service can be found in M5, those to improve its interoperability, user experience, and connectivity in M2.

Activity 2 – Depositing of data and curation of repositories: All Text+ Centers of TA4

providing repositories (UGOE/SUB, AWK-NRW) will contribute to Text+ Data Depositing by offering the opportunity to deposit research data in accordance with the respective documentation. Text+ Centers will pursue or renew certification processes for repositories (e.g. nestor, CoreTrustSeal) wherever not in place or expired, including re3data entries.

Activity 3 – Provision of contributed services: All Text+ Centers continue to provide contributed services (e.g. *research software directory, xml-archiving repository* (AWK-NRW), *lod.academy* (AWLM)**,** *ediarum, correspSearch, DTA base format* (all BBAW), *TextGrid* (UGOE/SUB)). Where support for contributed services is provided via the Text+ Helpdesk, the participants will allocate the necessary personnel to manage corresponding service queues.

In-kind contribution by partners (in PM): AWK-NRW (120), AWLM (12), BBAW (30), HAB (18), HAdW (6), h_da (6), MWS (10), NAWG (6), SAW (6), STI (24), UGOE/SUB (30), TUDa (30), KSW (3)

All achievements of M1 in TA4 correspond to the joint deliverables of M1 in 5.0.

TA4 M2: Interfaces between data from data domains

Co-Lead: AWK-NRW, AWLM

Contributors: BBAW, HAB, SAW, STI, TUDa

TA4 addresses interfaces between different catalogue data and finding aids, on the one hand, and between data sets on a content-related level, on the other. The aim is to make Text+ Registry entries available via other search systems and finding aids and vice versa (D4.2.1) as well as to enable cross-data-set and cross-domain searches on a content level (D4.2.2, D4.2.3, D4.2.4). NFDI consortia aim to aggregate interconnected data in shared spaces (such as Text+ Registry and FCS (SPARQL Endpoint - NFDI4Culture, n.d.), (Ritter, 2025), (NFDI4Objects, 2025)), enabling comprehensive cross-domain searches, reasoning, inference, and Retrieval-Augmented Generation (RAG). TA4 will contribute a range of integrable services on different levels of technical complexities from BEACON (2) to FCS (3, 4) with mapping strategies to increase cross-domain findability and interlinking. These services are not limited to editions by design but usable for various cross-domain research.

Activity 1 – Interfaces between catalogue data: The interoperability of the Text+ Registry will be further enhanced to establish its role as an aggregator and data hub in the NFDI (see also M1). This calls for continued curation and development of both the data model and interfaces in order to a) improve the automatic integration of metadata from external sources, and b) enrich data in existing edition catalogues by re-importing data that has been disambiguated, enhanced, and linked with authority files. A further aim is to integrate these curated data into library information systems and also to generally connect library systems more closely with the registry, so that the exchange is bilateral. In this regard,

critical measures to intensify collaboration internally with TA5 (see “Cooperation with other TAs”) and externally with DFG Specialised Information Services (SIS) as well as other domain experts have been established and will be continued and strengthened within phase II (see M5, and as proofs of concepts (*PhilFinder*, n.d.), (*nfdi.software*, n.d.)).

Activity 2 – BEACON Hub: One of the focal points is to promote the use of authority data, in particular but not delimited to interlinking with the GND (Text+ Consortium n.d.a). As a low-barrier offer, editions were supported in making their linked data available via BEACON-files (see (Al-Eryani et al., 2024); (Alvares Freire et al., n.d.); (Lordick & Hegel, n.d.)). As the *BEACON Findbuch* (Stephan et al., 2012), a central hub of the German BEACON network, has permanently ceased operations, Text+ will close that gap by deploying a new comprehensive BEACON hub at STI, in a technologically up to date form, i.e. as a knowledge graph (Voß & Schindler, 2017). Preparatory work and prototyping has been carried out by the consortium (Lordick, n.d.).

Activity 3 – Editions FCS: To increase the findability of editions as research data beyond the metadata-based approach, we will prepare the integration of editions into the FCS. This is closely linked to the question of derived text formats (see M4), which can be used as the starting point for a search interface to FCS and as result representation. While Text+ Centers will provide FCS endpoints (see M1), there is still a need to lower the barriers of integrating externally hosted editions by providing alternative possibilities to on-premise FCS endpoints in order to move the resources in the Text+ Data Space. To this end, we will compile typical use cases from the community to develop modules and provide corresponding processing services (transformation scenarios) as part of the FCS Integration Toolbox. The pre-processed data can then be transferred to Text+ Centers, which will integrate it into their search engines and associated FCS endpoints.

Activity 4 – Entity FCS: Using authority data makes the contents of the Text+ Data Space independent of domain, language, and spelling variants. The primary aim here is to identify the relevant entity types and the most frequently used authority data systems (GND, VIAF, WikiData, Geonames, etc.) as well as scientific knowledge bases so that they can be integrated into mechanisms such as automatic query expansion to other identifier systems on the basis of existing mappings. Furthermore, TA4 is collaborating with the research community to develop solutions for representing search results and designing user interfaces. This will address previously unsupported use cases and enable the most effective linking of resources. The aim is to develop a customised FCS Data View based on TEI-XML, for example, to support inline markup, which on the one hand functions as a derived text format via the API and on the other hand enables enhanced displays in the user interface of the search application.

In-kind contribution by partners (in PM): AWLM (24), BBAW (15), SAW (6), STI (6)
Requested funding: AWK-NRW (30), AWLM (30), HAB (15), SAW (18), STI (12), TUDa (30)

Table 5.25 Deliverables and milestones for TA4 M2

#	Month	Deliverable
M4.2.1	48	Text+ Registry for Editions linked to library catalogue system
D4.2.2	36	BEACON Hub: Deployment and publishing enhanced hub
D4.2.3	48	Editions FCS: Establish interfaces between editions and Text+ FCS
D4.2.4	60	Entity FCS: Deployment, mappings, customised FCS Data View

TA4 M3: Enabling text and language data for AI

Co-Lead: BBAW, TUDa

Contributors: AWK-NRW, JMU, STI, UGOE/SUB

TA4 contributes to the goals of the consortium by processing its textual data in formats suitable for mass processing by AI applications in digital editions and as training data (see M3.1) and by providing a forum to address needs of the community with regard to the application of AI-based tools.

AI and ML applications in scholarly editing are diverse and can be roughly categorized according to typical stages of the editorial workflow: digitization (e.g., OCR, HTR), pre-processing (e.g. layout recognition and segmentation), annotation (e.g. NER, POS-tagging, entity linking, and authority reconciliation), analysis (e.g. information extraction, content analysis, and further analytical functionalities), documentation, and publication (e.g. through code assistance). The advent of powerful and accessible LLMs has opened up new possibilities but also raised concerns and reservations around questions of data quality, transparency, traceability and accountability, and the introduction of biases in trained models (see M3 activity 2). Exchanges with the community on these issues form part of the M5 activities.

Activity 1 – Enhancing AI readiness of digital editions: Making digital editions accessible to AI applications (e.g. implementing chatbots) requires systematic transformation of editorial data into machine-readable formats and robust APIs for compliant access. This requires the development of conversion pipelines that transform XML-encoded texts into AI-optimized formats such as markdown while preserving structural information and, most importantly, maintaining licensing attribution and enabling copyright compliance (D4.3.1; see also M4).

Activity 2 – Bias assessment and quality assurance: Digital editions often reflect historical and editorial biases that can be amplified in machine learning models. Therefore, frameworks are needed to systematically evaluate representational biases across

demographics, periods, and genres within editorial corpora. Through comprehensive documentation of data set limitations and editorial methodologies, transparency about data constraints can be achieved. A set of guidelines of best practices for responsible AI use in editorial contexts will be established (D4.3.2 and D4.3.4), focusing on bias assessment and quality assurance.

Activity 3 – AI enhanced editorial toolchain: The integration of AI and ML into the editorial workflow can lead to more efficient and scalable digital edition production. A comprehensive AI-enhanced toolchain may combine multiple automated processes within a pipeline, from digitization and initial text capture, to data enrichment and linking, to entity recognition and mapping to authoritative sources (e.g. GND and WikiData). This complex pipeline can be further enhanced with automatic linguistic annotation, text structure recognition, and cross-reference detection, creating a continuous workflow from raw source material to enriched, structured editorial data (see also M4). TA4 will develop modular, interoperable, customizable tools to support the work of scholars in digital edition workflows respecting different textual traditions and editorial requirements but maintaining quality control through human oversight at critical decision points. In support of this goal, the guidelines for responsible AI use (D4.3.2 and D4.3.4) will include recommendations for chaining tools. Furthermore, we will identify suitable training data, models, and tools to accompany and enhance the editorial workflow and make them available to the community through the repository of reusable annotated corpora and pre-trained models and the curated software registry (Research Software Directory, n.d.) accordingly (D4.3.3).

Activity 4 – Tackling sustainability with AI: The community has long been concerned about the issue of digital editions becoming obsolete or disappearing entirely due to poorly maintained software (The Endings Project, 2025); (Holmes & Takeda, 2023), dubbed “Orphaned Editions”. With the productive use of ML and LLMs (e.g. co-piloting, code assistance), the cost of reviewing, software documentation, and refactoring web-based scholarly editions can be significantly reduced. In this scenario, TA4 will evaluate various procedures for creating future-proof digital editions, based on tried-and-tested software architectures for different levels of complexity, with a focus on AI-based migration for living systems (e.g., via TextGrid repository) and FCS as well as interface descriptions as part of data management plans. The aim is either to make orphaned editions executable again, in accordance with current community standards, or to convert them into formats that can be displayed using minimized technical frameworks (see (AGKI-DH, n.d.; I-D-E, 2025; DH-Kolloqium of the BBAW, n.d.)). Co-piloting can be performed on the basis of the Text+ Registry and LLMs provided by Text+ (see 5.5.3).

In-kind contribution by partners (in PM): -

Requested funding (in PM): AWK-NRW (15), BBAW (30), JMU (30), STI (6), UGOE/SUB (18), TUDa (30)

Table 5.26 Deliverables and milestones for TA4 M3

#	Month	Deliverable
D4.3.1	12	Conversion pipelines for editorial data
D4.3.2	18	Publication: Guidelines of best practices for responsible and sustainable AI use and development in editorial contexts (V1)
D4.3.3	12, 24, 36, 48, 60	Repository of reusable annotated corpora and pre-trained models and AI tools
D4.3.4	30	Enhancement of guidelines of best practices for responsible and sustainable AI use and development in editorial contexts (V2)
M4.3.5	36	Concept for tackling sustainability with AI

TA4 M4: Implicit and explicit structures of language

Lead: BBAW

Contributors: AWK-NRW, TUDa

Explicit structures are features of a text that are deliberately encoded to reflect specific research interests and scholarly priorities through their annotation or encoding format. These include the tagging of textual genesis and entities as well as other structured annotations along with decisions implied by the chosen format, whether hierarchically structured XML/TEI or a graph data model such as RDF. By contrast, implicit structures refer to properties and knowledge that remain hidden or undocumented but influence how a source is represented: First (relating to the text), properties of text and language which are not encoded such as part-of-speech structures. Second (relating to the edition as a whole), “tacit knowledge” based on methodological decisions, assumptions, and biases.

This distinction takes on particular significance in cases of multiple representations of the same source or of various derivative formats created from editorial data. Each additional manifestation makes different aspects explicit and introduces new layers of undocumented assumptions, thereby compounding the challenge of adequately describing what is encoded, what remains implicit, and how various manifestations of data relate to each other. During edition development, various data manifestations emerge (e.g. images, raw transcriptions, minimal tag sets). After publication, further manifestations arise through data reuse (e.g. normalized plain text with externalized metadata for quantitative analysis, markdown text for LLMs (see M3 activity 1), semantic triples for knowledge graphs (see M2)). Each manifestation embeds implicit editorial decisions about relevance and modeling, yet these assumptions often remain undocumented. This creates interconnected data sets that appear to represent the same source but actually reflect different scholarly perspectives, with relationships between them remaining opaque (Neuber et al., 2022a);

(Neuber et al., 2022b). TA4 addresses both aspects of explicit and implicit structures, making implicit knowledge visible and revealing connections between data manifestations with different explicit structures.

Activity 1 – Capturing implicit knowledge: a structured approach to data documentation and publication: While explicit structures are systematically documented through schemas and formal descriptions, implicit knowledge often remains inaccessible–frequently confined to brief remarks in editorial reports or retained as tacit knowledge. Without understanding the assumptions and processes that guided an edition's creation, researchers cannot fully assess data suitability for their research questions. TA4 will develop a framework for Implicit Knowledge Documentation, a structured guide that leads researchers through documenting their unspoken knowledge, decisions, and assumptions essential for data reuse. Based on this framework, TA4 will edit and publish data papers (e.g. in (ZfdG, n.d.) or (RIDE, n.d.)) that document both explicit and implicit aspects of data sets, making implicit knowledge accessible to the scholarly community while providing researchers with academic credit for their work.

Activity 2 – From implicit to explicit: revealing and managing multiple manifestations of data: The proliferation of interconnected yet isolated data sets that emerge from the same sources while explicating different research scopes creates significant challenges for data management. These data sets often exist in silos, making it difficult for researchers to leverage them effectively for cross-disciplinary or comparative studies. TA4 will evaluate several data formats for managing multiple data derivatives: METS for structuring and linking objects with metadata, PROV-O for transparent tracking of data transformation processes and derivation relationships, BagIt for ensuring data integrity through standardised packaging, and DataCite for enabling proper citation and persistent identification of data sets and versions. The outcome will be a conceptual framework with concrete technical guidelines and data models providing practical implementation strategies for managing complex data ecosystems in digital scholarly editions.

In-kind contribution by partners (in PM): AWK-NRW (60)

Requested funding (in PM): BBAW (30), TUDa (30)

Table 5.27 Deliverables and milestones for TA4 M4

#	Month	Deliverable
D4.4.1	24	Structured guide for documenting implicit knowledge
D4.4.2	40	Conceptual framework with concrete technical guidelines and data models for multiple manifestations of data
D4.4.3	50	Publication of peer-reviewed data papers (continuously until M60)

TA4 M5: Community activities

Co-Lead: AWK-NRW, TUDa

Contributors: AWLM, BBAW, BUW, HAB, JMU, UniRo, UniPB

TA4 will continue to serve as a hub for integrating, connecting, and networking the editions community with Text+ and its federated service infrastructure as well as the NFDI.

Activity 1 – Dissemination and outreach activities: Proven formats are further extended to include events and collaborations aimed specifically at previously underrepresented stakeholders (e.g. (REWISO+, n.d.), young researchers, etc.). With reference to the SIS network's action plan (FID-LG, 2022), concrete measures for cooperation have been agreed upon, regarding joint events, consulting and further outreach activities. For this, intensified collaborations with other consortia are integral, which will be achieved via NFDI sections to consolidate humanities consortia as a gateway for the community engaged with scholarly editions.

The following established formats will be continuously developed through activities by TA4, joint initiatives in Text+, and collaborations with external partners: “FAIR February”, dedicated to the FAIR principles, to be further developed with a focus on hands-on activities; “Edit-a-thons”, addressed to individual target groups in various formats and settings to continuously promote direct engagement and event-driven curation of the Registry; “FID/Text+ Jour Fixe”, jointly organised with SIS as a forum to initiate common activities tied to the working programme, is led by TA4 in the “AG FID Koop”; “Edit empfiehlt” (Edit empfiehlt – Text+ Blog, 2025), offering recommendations to honour creators of outstanding editions, and to provide orientation for editors and users. All events are accompanied by reports on their outcomes (e.g. publications, Text+ Blog, Webportal).

Engaging with previously underrepresented parts of the research community, e.g. young researchers, is to be achieved by participating and supporting a nationwide PhD Network (D4.5.1). The network is a continuation and extension of the cross-federal (Hessia, Rhineland-Palatinate) PhD network within the data competence center (HERMES, n.d.) and will be structured around regular online events focused on training and networking of doctoral candidates in cooperation with other humanities consortia of the NFDI. Additionally, workshops at conferences and other events (e.g. annual DHd conferences, Text+ Plenaries) on RDM topics (e.g. BEACON, FAIR RDM in DSE) will be held, including training on services and tools contributed by Text+ (e.g. (Ediarum, n.d.); (Dumont & Fechner, 2014); (BBAW, 2011)). Further joint activities beyond the immediate Text+ network are also planned (e.g. workshop series with the AG Germanistische Edition, renowned beginner and advanced schools on digital editing methods and technologies by IDE (I-D-E, 2025)).

Activity 2 – Training and education: The catalogue of existing tutorials relevant for working with digital editions (forthcoming), modelled in accordance with the DALIA

Interchange Format (Geiger et al., 2024), is to be maintained and expanded to integrate measures of quality control. Tutorials related to Text+ services will be continuously updated to ensure they reflect the current status of service development. Additionally, and in close cooperation with university partners of TA4, the selection of structured course module scenarios of a forthcoming catalogue of curricular recommendations will be expanded to include fundamentals of RDM teaching scenarios at bachelor and master levels. The "Guidelines for Quality Assessment and Assurance of Digital Editions" (Lemke et al., 2025) are to be maintained and updated.

Activity 3 – Consulting, helpdesk and support: After the successful establishment of Text+ Consulting, phase II will focus on the continuation, optimization and integration of the consulting service into NFDI-wide RDM-structures. To guarantee successful operation of Text+ Consulting and its helpdesk, TA4 will provide help desk agents to deal with user enquiries relating to domain-specific topics as well as regarding central services of Text+. Agents will also participate in open consultation hours (Text+ Research Rendezvous). Consultations are documented according to standards established in Text+ (*Text+ Consulting Guidelines*, 2025). One main advantage of these consolidation efforts consists in their connectability to the emerging RDM network of the NFDI section Education and Training (e.g. RDM knowledge base, NFDI wide helpdesk). This permits to draw on the expertise of the partner consortia as well as to contribute to enquiries from other consortia. TA4 coordinates the integration of Text+ Consulting to NFDI-wide structures as a major contribution to the consolidation and sustainability of the NFDI RDM infrastructure for our community of researchers.

In-kind contribution by partners (in PM): AWLM (24), BBAW (15), BUW (6), HAB (12), JMU (30), TUDa (60), UniRo (6), UniPB (15)

Requested funding (in PM): AWK-NRW (60), AWLM (30), HAB (15)

All achievements of M5 in TA4 correspond to the joint deliverables of M5 in 5.0.

Cooperation with other task areas

All TAs (see sections 5.1–5.4) share common measures with common objectives and deliverables (see 5.0). TA4 particularly cooperates with TA5 regarding a) 5.4.1 according to the further integration of TA4 into the Federated Data Infrastructure (see 5.5.1), b) 5.4.2, according to the linking of the Text+ Registry to library catalogue systems (see 5.5.1, activities 1 and 3), and c) 5.4.3, according to tackling sustainability with AI (see 5.5.3, activity 3). Regarding M5, a close coordination with TA5 and TA6 (Communication Office) is established concerning contributions to the Text+ Webportal (see 5.5.3, activity 2), to GOTRS for Text+ Consulting (see 5.5.3, activity 5), and the integration into NFDI wide RDM structures (see 5.5.3, activity 1).

Table 5.28 Risks of implementation TA4

Risk identified	Risk prevention and mitigation
Heterogeneity of the German library system regarding catalogue and discovery systems	Onboarding of VZG as Text+ co-applicant with Common Library Network (GBV) as partner, bringing in the K10plus union catalogue by VZG and BSZ; Close collaboration with Specialized Information Services (SIS) as infrastructure partners linked to libraries.
Copyright non-compliance and loss of scholarly provenance throughout the AI workflow	Development of conversion pipelines maintaining licensing attribution; implementing embedded metadata tracking in AI-based workflows.
Lack of adoption of services by editions community	Community approach to disseminate, collect feedback, optimise curating, data integration, and user experience.
Scalability of Text+ Consulting to manage high amount of complex RDM requests	Integration into the RDM network of NFDI Section EduTrain as the central NFDI hub of RDM expertise.

5.5 TA5: Infrastructure/Operations

Co-leads: HUB, UGOE/SUB, VZG

Contributors: AdWHH, DNB, GWDG, HSW, JSC, MWS, SAW, TUDD, UniBA

Infrastructure/Operations (IO) provides the technical framework and central services for the Text+ infrastructure, supporting language- and text-based research data in the humanities. We aim to provide users–our main stakeholder group–with a coherent, multi-faceted, and intuitive central access point to a constantly growing and rich Text+ Data Space to support them in the reuse, development, preservation, and machine processing of research data. TA5 pursues the following objectives:

1. Provide and maintain the Text+ portal as a user interface for Text+ services and resources, in collaboration with other task areas (TA5 M1, TA5 M3), for service and content deployment (objectives 1, 2, 4, 6, 7, 8).
2. Ensure a coherent Text+ service portfolio adhering to state-of-the-art standards and technologies (TA5 M1, TA5 M2, TA5 M4), including the provision of support resources for the data domains (TA5 M3), (objectives 2, 6, 7, 8, 9).
3. Provide LLMs as a basic service for data domain-specific use cases and service development (TA5 M4) (objectives 2, 3, 4).
4. Ensure adherence of Text+ services to user requirements and best practices in service development (TA5 M2, TA5 M4), (objectives: 1, 2, 3, 8, 9).
5. Support the connection and integration of other resources and services in the NFDI and beyond (e.g. ERICs, EOSC) (TA5 M1, TA5 M3, TA5 M4), (objectives 2, 7, 8, 9).

6. Provide means for a reliable and sustainable operation of the Text+ infrastructure (TA5 M4) (objective 9).
7. Provide a project collaboration environment for the consortium (TA5 M4), (objective 9).

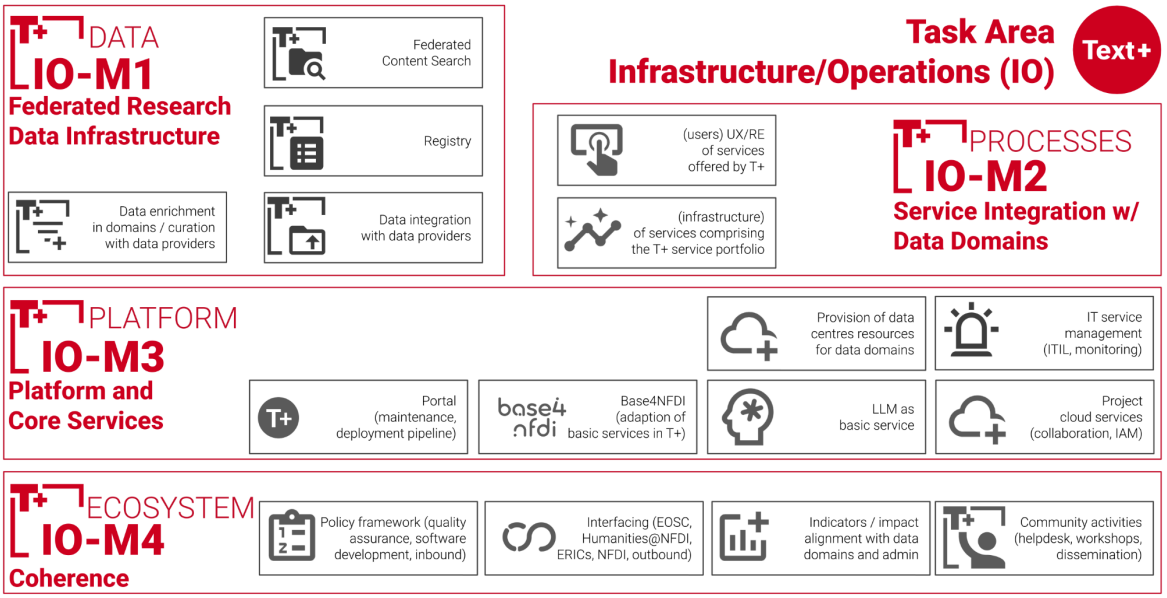


Figure 5.2: Measures of TA5 Infrastructure/Operations with associated activities

TA5 builds on the results of phase I and substantial preparatory work carried out by relevant national and European research infrastructures, in particular CLARIN, DARIAH, as well as Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V. as their common national hub and to the SSH Cluster of EOSC, notably the SSH Open Marketplace (SSH Open Marketplace, 2025). In addition to maintaining the existing portfolio of applications and services, the focus is on their expansion and targeted further development. In response to current developments in the scientific and technical fields, such as the increasing use of LLMs and the need for corresponding processing support, and to adapt continuously to a changing organizational and technical environment, such as the availability of new basic services as part of Base4NFDI (Bernard et al., 2023), new applications, services, and interfaces are also being developed for sustainable use. The target groups for all IO measures are:

- The Text+ (Data) Centers: TA5 supports activities around collection, storage, enrichment, and provision of data and services in an integrated research environment (primary responsibility for FAIRification of research data remains with data centers),

- the researchers of the scientific communities addressed by Text+: TA5 supports access to the Text+ data and service portfolio and advocates user interests within the consortium,
- projects in the national (NFDI) and international contexts (EOSC, CLARIN, DARIAH, SSH Open Marketplace) (Barbot et al., 2022; Rißler-Pipka & Weimer, 2023): TA5 will support the integration and linkage of data and services.

As a consequence, the TA's structure differs from that of the data domains and operates with the following four measures:

IO-M1 federated research data infrastructure focuses on the technical implementation of federated access to distributed research data sources. This includes the Federated Content Search (Körner & Eckart, forthcoming), Registry services (Genet et al., 2023a), and automated data integration pipelines. The measure also supports enrichment workflows through domain-specific curation in cooperation with external data providers. **IO-M2 service integration with data domains** facilitates interoperability and alignment of services within Text+ with specific domain-level requirements (Buddenbohm & Weimer, 2024). It covers both user experience and requirements engineering (UX/RE) as well as infrastructure-side service integration. The aim is to ensure that modular, reusable service components are effectively embedded into domain workflows. **IO-M3 platform and core services** provides the foundational infrastructure services for Text+, including: a central portal for coordination and maintenance of task area contributions, Base4NFDI components for standardized service provisioning, core services such as LLM-based tools, project-level cloud infrastructure, IT service management (ITIL-compliant), monitoring, and identity/access management (IAM). The measure ensures platform scalability, sustainability, and interoperability within the NFDI framework. **IO-M4 coherence** underpins strategic and operational coherence across the task areas. It defines policy frameworks for quality assurance, software development practices, and stakeholder interfacing (e.g., EOSC, ERICs, NFDI bodies). IO-M4 also manages performance indicators, impact alignment, and community-facing functions, including helpdesk services, dissemination activities, and workshop facilitation.

The TA plays an integrative role in relation to the service portfolio and infrastructure. In addition to focusing on the internal integration of the data domains, TA5 is also responsible for ensuring the connectivity of Text+ to the NFDI and other infrastructures such as ERICs (Buddenbohm, et al., 2021). TA5 addresses the measure structure of the data domains, the grand challenges, and the strategic themes of Text+, and builds upon the work carried out between 2021 and 2026. The Text+ architecture (Eckart et al., 2025) provides an infrastructural narrative and framing for the project. Two guiding principles are key to TA5's

collaboration with the other task areas (see also section 4.5 regarding the first and section 4.1 regarding the second principle and their relation to Text+ RDM strategy):

1. Maintaining a sustainable, secure, and effort-appropriate operational model,
2. thinking of infrastructure and services from the users' perspective.

Adherence to these two basic principles contributes substantially to the risk mitigation capacity of Text+. The central overview of the technical infrastructure and services, both in terms of internal and external use, allows the task area to develop quantitative metrics to monitor the overall (meta-)data quality and technical condition, the achievement of KPI-based project-specific targets, and the degree of utilization by the target groups. Unsustainable components, quality deficiencies or services that are not adopted by users will be identified by TA5, and possible solutions will be discussed with the data domains and/or project coordination.

IO-M1: Federated research data infrastructure

Lead: SAW

Contributors: AdWHH, DNB, GWDG, TUDD, UniBA, VZG

IO-M1 is closely related to the corresponding M1 in all data domains, and to a lesser extent to IO-M2, IO-M3, and IO-M4. It also relates to the core responsibilities of the Text+ Centers. The IO-M1 activities aim to support the Text+ Centers in their activities for the further development, utilisation, interlinkage, and dissemination of the federated Text+ Data Space and to ensure an integrated user experience for the data space on central applications and services. All activities of this measure are directly or indirectly related to these central components and are implemented in close coordination with the requirements of the data domains. Expanding and improving the usability, sustainability, and documentation of all IO services will improve the overall user experience and ensure long-term availability. IO-M1 includes the following activities:

Activity 1 – Search and retrieval infrastructure: This activity will further develop and provide the technical search & retrieval infrastructure for the federated Text+ Data Space and will also work on interlinking the existing tools to create an integrated search platform. The aim is to enhance the scalability, resilience, monitoring, and the overall Technology Readiness Levels of all services. For the FCS (Federated content search), the technology stack, user interface, software libraries, reference implementations, and web portal integration will be adapted to new data-domain-driven requirements. FCS hosting solutions, entity-oriented search based on GND and other authority files, and cross-data domain search functionality with integration of authority file APIs (incl. GND Reconciliation (Annisius et al., 2022) will also be supported. The activity will enhance the mapping, contextualization, and quality assurance of resources' metadata in the Registry. It will

provide a data contextualization service across multiple authority files and library catalogues (Schulz et al., 2025). The curation/search support for the Registry will also be expanded, e.g. by extending the resource/entity network & visualisation capabilities and other user-centred and API-based functionalities for search, exploration, and analysis.

Activity 2 – Data enrichment and curation: This activity includes the further development and provision of the environment and tool chains for data enrichment and curation to increase the interoperability of data in the Text+ Data Space and to improve the search and retrieval experience for users. It will provide and extend the data processing, integration, and enrichment workflows/pipelines and work on tools for the legally compliant provisioning of data (derived text formats), entity linking, and the preparation/preprocessing of data as AI training material. This includes continued work on the development of the GND Community Space (Kett et al., 2022; Buddenbohm et al., 2022a), which gives the Text+ community a low-threshold yet standardized way of cooperatively enriching the Text+ GND-based knowledge graph with entities and specialized, quality-assured information for the humanities, and on corresponding reference search applications and APIs that give access to this knowledge. The activity will be accompanied by intensive community work (workshops and training courses), supported by the GND agency Text+ (Fischer et al., 2024) which has been established during the first funding phase of Text+ and which is available as an operational service since 2024 (validation and conversion of data deliveries from stakeholders within the Text+ Consortium and beyond from other NFDIs, conversion of entityXML to MARC21 for integration into the GND, transfer of MARC21 data deliveries to the GND). EntityXML, a well-received outcome of phase I, will be deployed and operated (Sikora, 2022) for enriching research data with GND entities. Additionally, the Fluffy TextGrid Publication Workflow (Veentjer et al., 2025) will be extended to address requirements from the community, such as an optimised GUI, and improve code and quality of metadata to be published (e.g. creation of revisions, adding new content to published collections, import and publication of images including IIIF manifests). In conjunction with the import workflow, the TextGrid Repository portal will be enhanced with features such as an improved presentation and search functions (Calvo-Tello et al., 2023). LLM and NLP services, such as MONAPipe (Barth et al., 2023b) and the GWDG LLM resources, will be evaluated to identify and link entities in the metadata following the FAIR principles (Calvo-Tello et al., 2024b). This may serve as a template for import workflows at other Text+ Centers.

Activity 3 – Text+ Data Space expansion: This activity supports the expansion of the Text+ Data Space by integrating or linking it with other data sources (Mathiak et al., 2024), including data and metadata inventories and FCS endpoints, and by providing relevant consulting activities to achieve this. It will facilitate technical and content integration with

knowledge graphs within and beyond the NFDI context, enhance the use of authority data in resource descriptions, and support the interlinking of bibliographical data and research data by establishing API connections to library catalogues, in particular K10plus (Bibliotheksservice-Zentrum Baden-Württemberg, n.d.) as combined union catalogue of academic libraries in ten German states and the Prussian Cultural Heritage Foundation. The activity ensures the integration and compatibility with the European FCS infrastructure (CLARIN) (Körner et al., 2024) and supports the incorporation of international authority files, including Wikidata, into the search user interface. It organises workshops and hackathons, develops training materials, and provides comprehensive support for all services related to the core responsibilities of the Text+ Centers, such as Registry imports, FCS hosting solutions, and GND services.

In-kind contribution by partners: AdWHH (7), DNB (21,4), GWDG (30), UniBA (12), SAW (17), UGOE/SUB (15), VZG (24)

Requested funding: AdWHH (18), DNB (36), GWDG (30), UniBA (26), SAW (42), TUDD (24), VZG (26)

Table 5.29 Deliverables and milestones for IO-M1

#	Month	Deliverable
D5.1.1	21	Functional and online Search and Retrieval service within portal
D5.1.2	21	Functional data enrichment and curation pipelines
M5.1.3	24	Fully integrated search and retrieval infrastructure available (objectives: 2, 5), verified by D5.1.1, D5.1.2
D5.1.4	36	Error analysis for all search applications based on user feedback and failed user requests
D5.1.5	48	Data Domain focused training events regarding IO services related to Text+ Data Space (1 workshop/ hackathon per data domain (TA1-TA4)

IO-M2: Service integration with data domains

Lead: HUB

Contributors: UGOE/SUB, HSW, JSC, MWS, SAW, UniBA, VZG

IO-M2 relates to M2 of the data domains and, to a lesser extent, to M3. The aim is a) to improve the usability of all IO-related Text+ services through regular evaluation and iterative feedback, b) to integrate new or refined functionalities that are developed by the data domains, and c) to ensure that new services or functionalities adhere to the user experience and other technology standards set by the consortium. The central concern of the measure is to create a high level of coherence within the Text+ service portfolio, aiming to provide a simple, reliable, and consistent user experience. This involves adherence to uniform standards, clearly defined principles of use, and seamless linking of the individual components based on the underlying Text+ architecture. The coordinated and modelled

interaction and implementation of the various service components increases accessibility for a wide range of users. For this, a structured process for collecting, evaluating, and integrating user feedback is required to ensure that the Text+ services meet user requirements and provide a good user experience for the target groups. The following activities pursue this:

Activity 1 – User experience: This activity makes manifest consistent reliability and usability across the Text+ service portfolio by standardizing functionalities, interactivity, and interfaces. It will develop and implement a structured process for regularly gathering user feedback, evaluating services, and providing actionable feedback to service providers. It regularly applies and publicly documents the user requirements (UR) and user experience (UX) indicator set developed in IO-M4. Guidance and support–both in terms of guidelines, design templates, but also in regular evaluation feedback reports and public result documentation (Föhr et al., 2023)–are provided to the Text+ services for analyzing user requirements and implementing a uniform user experience across the Text+ portfolio. Together with the other task areas, integration strategies are continuously refined based on identified usability issues and identified deficits.

Activity 2 – Integration of search and retrieval infrastructure: In phase I, FCS and Registry were developed separately as part of the overall search & retrieval infrastructure of the Text+ service portfolio. While TA5 IO-M1 will provide the technological basis for development, this activity will support domain-specific enhancements for both services by ensuring their compatibility with the overall service. It will also improve the interlinking between the two services to achieve a more seamless user experience.

Activity 3 – Workflow Integration (IO and data domain-driven): This activity identifies relevant scientific workflows that could be integrated to enhance the Text+ service portfolio in the data domains, like e.g. machine learning workflows (Sherpa et al., 2024). It will improve user experience by effectively interlinking processing tools, analysis tools, core services, and data domain-specific services (see Activity 4). Additionally, it warrants seamless integration among IO services.

Activity 4 – Service integration (data domain-driven): This activity assists data domains in integrating their specific services into the Text+ service portfolio (Ďurčo et al., 2023; Buddenbohm & Weimer, 2024), adhering to consistent quality standards related to UX, UR, and service design. It prioritizes the principle “integration before implementation”, emphasizing the integration of existing sustainable services (such as those from Base4NFDI projects) before creating new services within Text+ (refer to section 4.4). A priority is to facilitate connections to the core services (TA5 IO-M3).

In-kind contribution by partners: MWS (6), HSW, HUB (15), SAW (16), UGOE/SUB (30), UniBA (12), VZG (20)

Requested funding: MWS (6), HUB (30), SAW (42), UGOE/SUB (30), UniBA (26), VZG (20)

Table 5.30 Deliverables and milestones for IO-M2

#	Month	Deliverable
D5.2.1	21	UR/UX indicator report for selected Text+ services
D5.2.2	21	FCS / Registry integration report
M5.2.3	24	Integrated resources, including concise attribution to data centers, provided [objectives: 1, 2, 9], verified by D5.2.1 and D5.2.2
D5.2.4	57	Modelled and tested UX evaluation and feedback framework

IO-M3: Platform and core services

Lead: GWDG

Contributors: JSC, UGOE/SUB, TUDD, VZG

TA5 IO-M3 focuses on the sustainable operation, coordinated development, and strategic integration of services into the Text+ infrastructure to ensure seamless user access and support for research communities. It links to all data domain measures as it provides basic services and acts as the presentation layer for the resources and services offered to users. It pursues this goal through the following activities:

Activity 1 – Integration of Base4NFDI services: Current integration activities of Text+ include: IAM4NFDI, PID4NFDI (Böhm, 2024), Jupyter4NFDI, DMP4NFDI, nfdi.software, and RDMTraining4NFDI (see also 3.2.3). Upcoming features and additional services are integrated according to the RDM Strategy of Text+ (see 4.4).

Activity 2 – Portal management: Operates the Text+ portal as the central access point for the community. This includes supporting all task areas in web content management and coordinating the deployment pipeline for content updates. It does not include content creation and integration. This responsibility remains with TA Administration as the authoritative body in cooperation with data domains' M5 community activities.

Activity 3 – Provision of LLM resources for research scenarios: to support domain-specific, research-driven use cases. The development and implementation of LLM-based services–such as prompting assistance, data curation, or working with orphaned editions–are driven by the domain-specific task areas.

Activity 4 – Provision of data center resources for data domains enables an ad-hoc provisioning of requested resources by the data centers for data domain specific activities. This may include processing or storage capacities or support in development and deployment activities. Also includes tools and server capacity for data management and processing and high performance search engines. The archive for long-term preservation (Dogaru, 2023), which has been developed during phase I is operated through this activity .

Activity 5 – Provision of project cloud services facilitates collaboration and communication across the consortium, as this activity provides and maintains, in continuation of the first funding phase (Buddenbohm et al., 2022b), cloud-based project services, including: Nextcloud (file sharing); Matrix (messaging); Indico (event management); GOTRS (Friedrichs & Jander, 2022); OpenProject (project management); collaborative pads; mailing lists; and general other IT infrastructure. Additional services may be deployed. Some services are also available–and have already been deployed–for other consortia, such as the helpdesk for KonsortSWD and the Matrix instant messaging service.

Activity 6 – IT service management and technical monitoring: ensures the reliable operation of the entire Text+ infrastructure through proactive IT service management. A public-facing technical monitoring dashboard provides real-time insights into the operational status of all services. Activity aligns in this regard with GWDG's DIN EN ISO/IEC 27001.

In-kind contribution by partners: GWDG (30), UGOE/SUB (15), TUDD (3), VZG (13)

Requested funding: UGOE/SUB (30), GWDG (30), TUDD (6), VZG (10)

Table 5.31 Deliverables and milestones for IO-M3

#	Month	Deliverable
D5.3.1	21	Provisioning of data center resources for data domains, ensuring proper attribution to data centers
D5.3.2	24	Provisioning of project cloud services for task areas, documentation through poster
M5.3.3	54	Text+ portal including content and service deployments fully operational [objectives: 1, 6, 7, 8]
D5.3.4	60	Integration of base services and documentation through news and blog items

IO-M4: Coherence

Lead: UGOE/SUB

Contributors: HUB, MWS, DNB

IO-M4 ensures that the project achieves its goal of building an integrated research data infrastructure on time, within scope, and in line with established standards. The infrastructure must fulfill FAIR principles as well as support tools and services for research data management and integration with other NFDI consortia. At the same time, it must be agile and sustainably operable to meet future demands (e.g. "One NFDI", EOSC node). IO-M4 is a crucial risk mitigation instrument for IO and the project as a whole (see the risk mitigation table at the end of this chapter). This overarching aim will be implemented as a practical framework that supports both data domains (research-driven needs) and infrastructure partners (operational sustainability) through the following activities:

Activity 1 – Community activities: Promote the use, development, and awareness of the Text+ portfolio efficiently. Target groups include internal stakeholders (data domains, data

centers) and individual users, as well as other infrastructures (especially ERICs, Humanities@NFDI, EOSC, NFDI). This dialogue will be facilitated with an annual event “State of the Infrastructure”.

Activity 2 – Policy frameworks: Concern inbound coherence and interoperability regarding quality assurance, software, and service development. The principle is “only as much regulation as necessary”, including documenting service policies, Text+ standards, software development guidelines, and quality assurance (Weimer & Buddenbohm, 2023). This measure relates to the respective M5 Community Activities of the data domains and TA Administration.

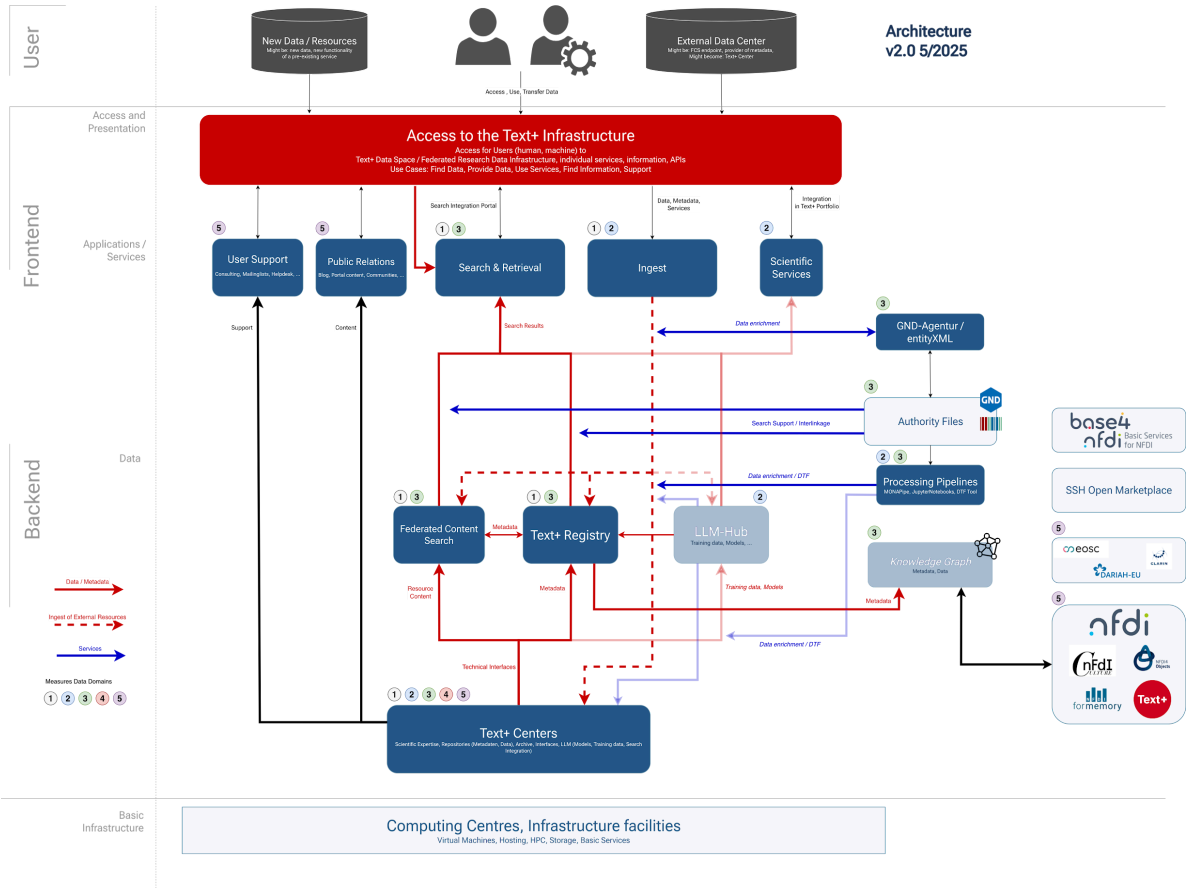


Figure 5.3: Text+ Architecture [version 2.0, 5/2025]

Activity 3 – Indicators/impact alignment: Involves continuous communication with other task areas to ensure Text+ offerings align with user requirements and the overall consortium goals in close collaboration with TA Administration for project-wide work and results monitoring/DoW compliance when it comes to infrastructure and services.

Activity 4 – The Text+ Architecture (Eckart et al., 2025) will be continuously refined and serve as the main instrument for facilitating the discussion with all stakeholders, particularly the data domains on service portfolio, infrastructure, interoperability, sustainability, and use aspects. IO-M4 relates to M5 of the data domains in the sense that it ensures consideration

and adaptation of user requirements. Apart from this, IO-M4 maintains a strong link to TA6 in terms of sustaining an up to date infrastructure and service portfolio.

In-kind contribution by partners: DNB (12), HUB (30), MWS (12), UGOE/SUB (30)

Requested funding: HUB (30), MWS (12), UGOE/SUB (30)

Table 5.32 Deliverables and milestones for IO-M4

#	Month	Deliverable
D5.4.1	12	Quality assurance of Text+ resources and services: provision of an indicator set and handed over to TA5 M2 after deployment
D5.4.2	24	Guideline on services policy and software development published
M5.4.3	60	Text+ service portfolio according to unified quality standards fully operational [objectives: 3, 7, 8], verified by D5.3.1, D5.3.4
D5.4.4	60	Text+ Architecture 2.0 in relation to NFDI Overall Architecture & the EOSC node concept
D5.4.5	12, 24, 36, 48, 60	Event: State of the Infrastructure

Cooperation with other task areas

The data domains and therefore the respective task areas (see sections 5.1–5.4) share common measures with common objectives and deliverables. The respective details are described in the introduction of the Work Programme (see section 5.0). IO targets all data domains (TA1–TA4) with its measures and also forms the interface to the NFDI for infrastructural and service aspects, particularly for basic service developments. It also works closely with TA6 Administration on overarching NFDI activities and cross-cutting topics. IO contributes to an integrated NFDI structure (OneNFDI) (Buddenbohm & Steckel, 2024) with domain-specific services and data offerings.

Table 5.33 Risks of implementation TA5

Risk identified	Risk prevention and mitigation
Over-complicated service integration	Develop and maintain clear, well-documented integration guidelines. Foster close collaboration with data domains to align services early in development. Apply iterative testing and feedback loops to ensure seamless integration (see Measure 2).
Delay of Base4NFDI services	Engage as a partner or closely collaborate with Base4NFDI service developments (see section 3.2). Monitor progress through TA5 M4 and trigger preventing/mitigating decisions early in the process.
Lack of user adoption	Continuously collect user feedback through surveys, focus groups, and direct interaction. Integrate this feedback into the development cycle to improve usability. Provide training and support to encourage engagement (see Measure 2 and 4).
Technical and system failures	Implement robust IT service management (ITIL-compliant) protocols. Install redundant systems and backup solutions are in place. Monitor services actively using real-time monitoring dashboards and regularly conduct system health checks (see Measure 3).
Insufficient scalability	Design the infrastructure with scalability in mind from the outset. Utilize cloud-based services and modular components that can be easily expanded. Monitor service

	performance and adjust resources as needed (see Measure 4, 3, 2).
Security and compliance issues	Adhere to strict data security protocols, including encryption, access control, and regular security audits. Ensure compliance with relevant laws (e.g. GDPR) and regularly update policies (see Measure 4).
Resource allocation conflicts	Establish clear roles, responsibilities, and communication channels. Implement a project management system for tracking resource usage and priority allocation. Review progress regularly and adjust resources as needed (see Measure 4 and task area administration).
Insufficient collaboration with external infrastructures	Regularly participate in cross-infrastructure working groups and maintain communication with relevant partners. Monitor compliance with shared standards and protocols for interoperability (see Measure 4).
Ineffective documentation	Develop comprehensive and user-friendly documentation for all services and tools. Ensure that documentation is continuously updated and easily accessible. Implement a dedicated team for ongoing documentation review and improvements (see Measure 4).
Sustainability challenges beyond project duration	Work closely with the NFDI and partner institutions to secure funding and operational support beyond the project's initial phase. Develop a clear sustainability plan and explore potential revenue models or collaborations for continued support (see Measure 4 and task area administration).
Technical debt	Follow best practices in software development, including modular design and regular code reviews. Allocate time and resources for addressing technical debt and refactoring code as necessary (see Measure 4).

5.6 TA6: Administration

Co-leads: UGOE/SUB, IDS (Note: All measures in this task area will be worked on jointly by the two partners, IDS and SUB. Therefore no further information on co-leads and contributors is provided.)

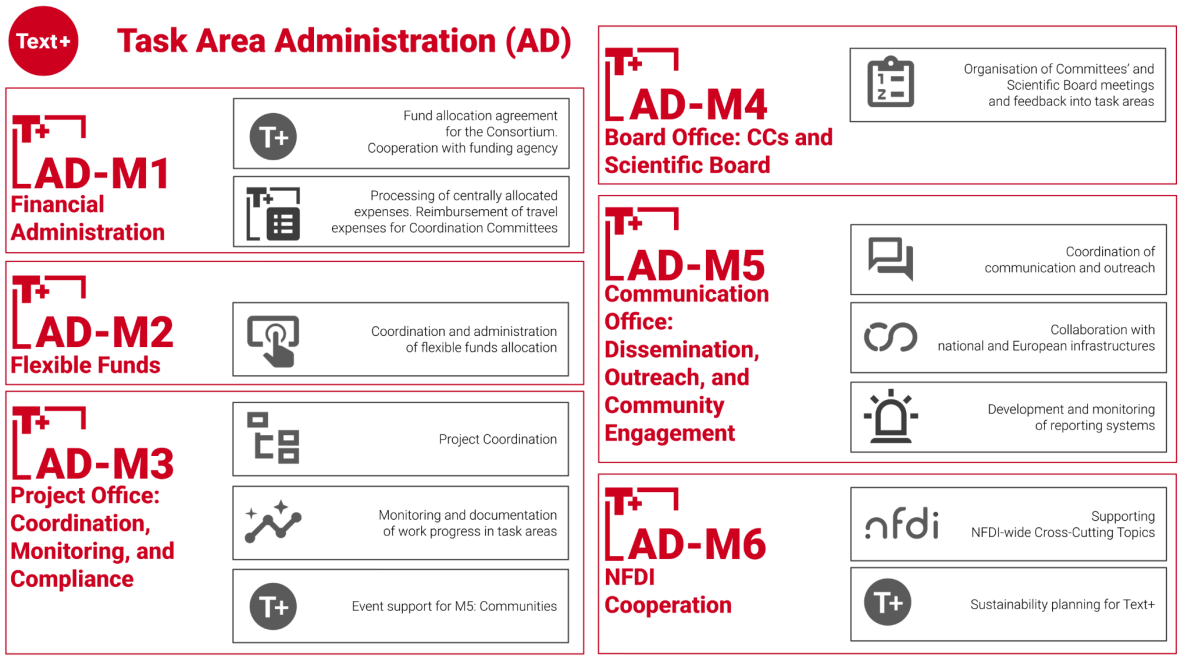


Figure 5.4: Measures of TA6 Administration with associated activities

The TA Administration coordinates all task areas so that they will deliver an integrated research data infrastructure. In close collaboration with IO (Infrastructure/Operations) and

the Text+ Centers (M1 in all data domains), it emphasizes cooperation and interoperability within the NFDI (NFDI association and office, all NFDI consortia, and Base4NFDI), with other research infrastructures (e.g., ERICs, EOSC), and with the user community. TA Administration promotes agile, participatory, and solution-oriented working methods, enabling the joint development of services and flexible responses to emerging needs. It is divided into the following measures, which align with the NFDI-wide coordinated recommendations on the establishment of consortial governance.

AD-M1: Financial administration of the consortium

Activity 1 – Fund allocation agreement: This measure involves the administration of the cooperation and fund allocation agreement for the consortium.

Activity 2 – Processing of centrally allocated expenses: It involves the disbursement of project funds and flexible funds (see Measure 2) to co-applicant and participant institutions. The package also covers the potential processing of funds from Base4NFDI if Text+ partners are involved in basic service requests. Continuous monitoring of the total annual budget and the expenditures of individual co-applicants and participant institutions is part of the responsibilities, along with preparing and supporting the annual expenditure reports. The work package coordinates financial transactions with the funding agency as well as financial reporting to the agency. Relating to these topics, there is also an exchange on financial issues with NFDI bodies (e.g., NFDI-Finanzkreis). AD-M1 handles the reimbursement of travel expenses for Coordination Committee members and processes centrally allocated expenses for plenaries, additional workshops of the task areas, and travel costs for central functions.

In-kind contribution by partners (in PM): UGOE/SUB (15), IDS (6)

Requested funding (in PM): UGOE/SUB (75)

Table 5.34 Deliverables and milestones for AD-M1

#	Month	Deliverable
M6.1.1	6	Adapted reporting chain from partners to central financial administration implemented
D6.1.2	12, 24, 36, 48, 60	Annual report on financial situation of consortium, including disbursement of project funds to partners, reimbursement of travel costs of CC members, occasion-driven processing of cash flows

AD-M2: Flexible funds

Activity 1 – Coordination and administration of flexible fund allocation: The Text+ budget includes flexible funds applicable for extensions of the portfolio of data, services, and further Text+ related activities. It follows a documented procedure aiming at enhancing and extending Text+ service and infrastructure offerings as well as community activities. M2

coordinates with M1. The application process will begin already during the first funding phase by the existing governance and then be handed over.

In-kind contribution by partners (in PM): -

Requested funding: UGOE/SUB (15 PM), flexible funds (2.230.000€)

Table 5.35 Deliverables and milestones for AD-M2

#	Month	Deliverable
M6.2.1	3	Annual application procedure for flexible funds is re-modelled (including reporting loop to TA6 M1)
D6.2.2	12, 24, 36, 48, 60	Report on disbursement of flexible funds

AD-M3: Project office

Activity 1 – Project coordination: The office acts as the central service and coordination unit of the consortium, supporting leadership in governance, external representation, and partner communication as well as all governance structures (see 3.4). It ensures smooth operations by liaising with national (e.g., NFDI task forces and groups) and European (e.g., ERICs, EOSC) infrastructure initiatives.

Activity 2 – Monitoring and documentation of work progress: A core responsibility is to monitor and document progress across all task areas, focusing on the quality, usage, and visibility of Text+ services and data (incl. reporting to the DFG). The office ensures that all partners have access to key services and information.

Activity 3 – Event support: It warrants that workshops, training sessions, public consultations, and annual consortium meetings are well aligned with project objectives through an agile and transparent decision-making process.

In-kind contribution by partners (in PM): IDS (15), UGOE/SUB (30)

Requested funding (in PM): IDS (30), UGOE/SUB (30)

Table 5.36 Deliverables and milestones for AD-M3

#	Month	Deliverable
D6.3.1	6	Set up of a regular reporting chain from task areas to TA6 resulting in regular bulletins encompassing all project governance elements
D6.3.2	6	Establishment of a process for organizing annual consortium meetings, hosted by different partners each year with organizational support by the office
D6.3.3	6	Set up of a process of coordination and cooperation with the steering group, preparation and reporting of steering group and management group meetings
M6.3.4	12	New fully integrated feedback and reporting chain from task areas to TA6 is implemented for progress tracking in task areas
D6.3.5	12, 24, 36, 48	Evaluation of project governance's performance and risk mitigation together with spokesperson(s)

D6.3.6	12, 24, 36, 48, 60	Annual report based on a regular activity report (on helpdesk activities, liaison activities with national and international infrastructure initiatives, etc.)

AD-M4: Board office

Activity 1 – Organisation of committees' and scientific board meetings and their communication: This measure supports all coordination committees and their chairs and the scientific board by organizing meetings, collecting progress reports, assisting with flexible fund processes, and documenting outcomes. Meetings may be held on a per-committee basis as well as for bringing all committees together in joint sessions.

In-kind contribution by partners (in PM): UGOE/SUB (15), IDS (15)

Requested funding (in PM): IDS (35), UGOE/SUB (15)

Table 5.37 Deliverables and milestones for AD-M4

#	Month	Deliverable
D6.4.1	6	Roadmap for the coordination committees and the scientific board
M6.4.2	6	Working environment, processes, and tasks for boards and committees are communicated
D6.4.3	12, 24, 36, 48, 60	Preparation and reporting of CC meetings, administrative support of CC chairs

AD-M5: Communication office

Activity 1 – Coordination of communication and outreach: This measure aims to coordinate communication and outreach activities across all task areas within Text+. It will manage communication channels, coordinate the web portal content developed in the task areas, oversee the helpdesk, and establish a news desk with representatives from the task areas. A communication concept will be used to align the TA1-4 M5 activities and pay regard to the coherence efforts in IO-M4.

Activity 2 – Collaboration with national and European infrastructures: The package ensures close collaboration with national (e.g., NFDI, research data management) and European (e.g., ERICs, EOSC) infrastructures, developing adaptable content tailored to diverse audiences.

Activity 3 – Development and monitoring of reporting systems: Based on the communication concept, a reporting system tracks KPIs such as publications, talks, and event contributions in collaboration with IO-M4 and TA 1-4 M1.

In-kind contribution by partners (in PM): -

Requested funding (in PM): IDS (30), UGOE/SUB (30)

Table 5.38 Deliverables and milestones for AD-M5

#	Month	Deliverable
D6.5.1	3	Provision of a project bibliography for all task areas
M6.5.2	12	Participatory design of a strategic communication and outreach concept for the consortium, including relevant external actors (e.g. Humanities@NFDI, Base4NFDI, NFDI Jour fixe communication)
D6.5.3	12, 24, 36, 48, 60	Concept of and annual report based on a regular activity report of the task areas (on outreach activities, training activities, achievements)

AD-M6: NFDI cooperation and sustainability

Activity 1 – Supporting NFDI-wide cross-cutting topics: This activity supports Text+ partners in contributing to NFDI-wide cross-cutting topics by coordinating input, organizing working group meetings, and assisting Text+ experts in administrative tasks. It fosters cooperation with the NFDI sections, Base4NFDI, particularly with the Humanities@NFDI consortia, as well as bilateral cooperation with other consortia, including those (such as KonsortSWD) for which a Memorandum of Understanding already exists.

Activity 2 – Sustainability planning: It also coordinates sustainability planning for Text+, including updating governance documents, preparing a new consortium agreement, and contributing to NFDI-wide sustainability efforts beyond the current funding period.

In-kind contribution by partners (in PM): UGOE/SUB (30), IDS (24)

Requested funding (in PM): IDS (25), UGOE/SUB (15)

Table 5.39 Deliverables and milestones for AD-M6

#	Month	Deliverable
D6.6.1	12, 24, 36, 48, 60	Report on NFDI cooperation
M6.6.2	30	Text+ is an active partner in cross-NFDI collaboration with other NFDI consortia and NFDI bodies; national and international collaborations exist.

Cooperation with other task areas

The TA Administration is central to the progress and success of the overall project. It monitors proper cooperation between all project bodies and ensures a motivating, result-oriented, and cooperative communication and working mode. Furthermore, it supports the spokesperson and functions as a communication relay to the community, the NFDI, and other external entities (such as ERICs, EOSC).

Table 5.40 Risks of implementation TA6)

Risk identified	Risk prevention and mitigation
Delays in Financial Reporting and Fund Allocation	Establish clear internal processes for reporting and financial monitoring for timely submission. Implement an automated reporting system for tracking and streamlining financial documentation. Conduct regular audits, including possibly check-ins, with partner institutions to monitor their expenditures. (Measure 1).

Lack of Transparency in Flexible Funds Allocation	Develop a transparent, competitive procedure for allocating flexible funds, building on phase I practices and adapting them as needed. Institute inclusive design, equal access, and regular monitoring to maintain transparency.
Ineffective Feedback Loops	Establish structured and formal feedback mechanisms to gather input from all task areas. Regularly evaluate and refine feedback processes to ensure they are timely and actionable. Schedule recurring meetings to audit the implementation of changes based on feedback. Monthly report to the meeting of the steering group.
Insufficient Dissemination and Outreach	Develop a comprehensive dissemination and outreach strategy with a clear reporting mechanism. Ensure coordination with all task areas to maintain consistent messaging and maximize visibility. Monitor the impact of dissemination activities and adapt strategies accordingly.
Insufficient Coordination Across Task Areas	Establish a clear and structured process for regular communication between all task areas. Set up frequent cross-task area meetings to foster alignment and survey consistency in activities for all overarching measures, coordinated with the responsible Co-Spokesperson. Monthly report to the steering group. Provision of short version of steering group minutes to the consortium, including the reports on task area progress and cross-cutting activities, dates of working group meetings.
Challenges in NFDI Cooperation and Sustainability	Develop a roadmap for NFDI cooperation with clear timelines and responsibilities. Regularly engage in discussions with other consortia (e.g., NFDI-Finanzkreis, NFDI-Kommunikationskreis, NFDI-Managementkreis, NFDI-Sprecherinnenkreis), sections of NFDI, NFDI working groups, and the NFDI Directorate to ensure alignment. Monitor the long-term sustainability of the project and update governance documents as needed.

6. Additional aspects

6.1 Equity and diversity

Text+ is committed to equity and diversity in all dimensions–disciplinary background, gender, age, career stage, and institutional affiliation. This commitment is reflected in both governance structures and operational practices.

Governance and representation: In the first project phase, Text+ faced an imbalance in its steering group in several dimensions. Most seats, and ultimately all of them, are held by men, with a majority of affiliations with non-university research institutions. With the reorganization for the second phase, a more balanced representation has been achieved: Co-applicant institutions map to a range of universities, non-university research institutions, and infrastructure providers. With six female and five male co-spokespersons gender balance is reached for the steering group, including a broader age and career stage diversity, and each task area is co-lead by a mixed team in terms of gender and institutional affiliation.

To further improve equity and diversity aspects, bylaws and procedures of each governance group will be revised and include provisions to prevent dominance by one group – be it by gender, institutional type, or task area. While legal constraints across institutions pose challenges, Text+ is committed to implementing these changes wherever possible.

Participatory processes: In all elections (e.g. for coordination committees), gender balance is explicitly addressed. Additional candidates may be nominated to ensure equal representation. The scientific board already reflects a balanced composition, and this principle continues to guide the formation of all committees.

Events and visibility: Programme committees are instructed to ensure gender balance among invited speakers and contributors. This has been successfully implemented in past events, with the exception of presentations by (co-)spokespersons.

Disciplinary diversity: Text+ brings together researchers from a wide range of disciplines, as defined by the DFG classification (2024–2028), including philology, linguistics, literary studies, philosophy, political science, musicology, and computer science. This diversity ensures that the consortium addresses the needs of the full spectrum of humanities disciplines working with text- and language-based data. This diversity also warrants that Text+ effectively addresses the needs of humanities disciplines working with text- and language-related data.

6.2 Declaration of generative AI and AI-assisted technologies in the writing process

The authors acknowledge the use of Large Language Models (LLMs) as writing aids in phrasing this application, based on the authors' notes, ideas, and concepts, especially for grammar, spelling, and style. The consortium retains full responsibility for the content.

NOTE

Please note: the financial tables and personal contributions, which are included in the full version of the application, are not included in the community version.

Also not included are Appendices 2 (CVs of the Co-Spokespersons) and 3 (Letters of Commitment).

Appendix 1: Bibliography and list of references

Items (co-)authored or (co-)developed by members of the consortium are set in bold.

AGATE – A European Gateway for the Academies of Sciences and Humanities. (n.d.). Retrieved 1 July 2025, from https://agate.academy/

***AGKI-DH DHd-AG Angewandte Generative KI in den Digitalen Geisteswissenschaften*. (n.d.). Retrieved 1 July 2025, from https://agki-dh.github.io/**

Al-Eryani, S., Gerstner, E.-M., Hegel, P., Lordick, H., Schnöpf, M., Schulz, D., & Sikora, U. (2024). *BEACONe unde venis et quo vadis?* https://doi.org/10.5281/ZENODO.10732655

Albertoni, R., Browning, D., Cox, S. J. D., Gonzalez-Beltran, A., Perego, A., & Winstanley, P. (2024, August 22). *Data Catalog Vocabulary (DCAT)–Version 3. W3C Recommendation*. https://www.w3.org/TR/2024/REC-vocab-dcat-3-20240822/

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2024). Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, *8*(1), 17. https://doi.org/10.1007/s42001-024-00345-9

Alvares Freire, F., Lordick, H., & Hegel, P. (2025). *FAIR February 2025: Veranstaltungsreihe*. I wie BEACON. https://events.gwdg.de/event/948/page/197-19022025-i-wie-beacon

Annisius, M., Fischer, B. K., & Steckel, A. (2022, October 25). *Ein Baum in der GND*. https://doi.org/10.5281/zenodo.7249018

Barbot, L., Gray, E., Fischer, F., Ďurčo, M., König, A., Puren, M., Buddenbohm, S., Concordia, C., & Illmayer, K. (2022, May 25). *The SSH Open Marketplace: A multi-voiced story*. https://doi.org/10.5281/zenodo.6580303

Barth, F., Blätte, A., Calvo Tello, J., Debbeler, A., Draxler, C., Fischer, S., Fladrich, M., Funk, S. E., Genêt, P., Göbel, M., Hemmer, A., Hug, M., Knappen, J., Krielke, M.-P., Kurzawe, D., Lehmberg, T., Leinen, P., Mollenhauer, E., Rau, F., … Zinn, C. (2023a). *Leitlinie für das Integrieren von Daten in Text+/NFDI*. https://doi.org/10.5281/zenodo.12744056

Barth, F., Bracke, Y., Calvo Tello, J., Dogaru, G., Dönicke, T., Du, K., Funk, S. E., Genet, P., Göbel, M., Keller, L., Kurzawe, D., Veentjer, U., & Weimer, L. (2023b, October 10). *MONAPipe: Modular Natural Language Processing Pipeline for Digital Humanities*. https://doi.org/10.5281/zenodo.8424925

BBAW (Ed.). (2011, 2025). *Deutsches Textarchiv – Basisformat*. https://www.deutschestextarchiv.de/doku/basisformat/

Bernard, L., Altenhöner, R., Böhm, F., Diepenbroek, M., Ebert, B., Fluck, J., Herres-Pawlis, S., Klinger, A., Koepler, O., Lorenz, S., Mathiak, B., Miller, B., Pelz, P., Rißler-Pipka, N., Ritz, R., Sax, U., Schimmler, S., Schörner, T., Schrade, T., … Suchodoletz, D. von. (2023). *Base4NFDI - Basic Services for NFDI*. https://doi.org/10.5281/zenodo.10245518

Bibliotheksservice-Zentrum Baden-Württemberg. (n.d.). *BSZ-GBV-Wiki zu K10plus*. Retrieved 25 July 2025, from https://wiki.k10plus.de/

Böhm, J. (2024). *Use Case Analysis: PID Adoption in the Text+ Consortium using the Example of SUB Göttingen*. https://doi.org/10.5281/zenodo.14327691

Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., Jong, F. de, Kamocki, P., König, A., Lindén, K., Navarretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadiņa, I., Trippel, T., Witt, A., & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09658-z

Brünger-Weilandt, S., Bruhn, K.-C., Busch, A. W., Hinrichs, E., Maier, G., Paulmann, J., Rapp, A., von Rummel, P., Schlotheuber, E., Schmidt, D., Schrade, T., Simon, H., Stein, R., & Teich, E. (2020). *Memorandum of Understanding by NFDI Initiatives from the Humanities and Cultural Studies*. Zenodo. https://doi.org/10.5281/zenodo.4045000

Buddenbohm, S., Jander, M. L., Trippel, T., Weimer, L., & Walker, N. (2021). *In-kind Contributions for two ERICs in one National Initiative: Practices and Experiences in CLARIAH-DE*. Zenodo. https://doi.org/10.5281/zenodo.5516816

Buddenbohm, S., Fischer, B., Kett, J., Mayer, D., Stein, R., & Trippel, T. (2022a, June 23). *There are a million ways to describe the world–How to share your authority without losing control*. https://doi.org/10.5281/zenodo.6683365

Buddenbohm, S., Weimer, L., & Wienert, A. (2022b, December 9). *Werkzeuge zur Kommunikation, Kollaboration und Arbeitsorganisation in Text+*. Text+ Plenary 2022, Mannheim. https://doi.org/10.5281/zenodo.7229455

Buddenbohm, S., & Steckel, A. (2024, September 5). *Systematischer Umgang mit Forschungsdaten am Campus Göttingen*. https://doi.org/10.5281/zenodo.13693877

Buddenbohm, S., & Weimer, L. (2024). *Text+ Services Policy* [Report]. Göttingen State and University Library. https://shs.hal.science/halshs-04698855

Bundesministerium für Bildung und Forschung. (2024, February 28). ***Eckpunktepapier zum Forschungsdatengesetz*****.** **https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/sonstige/Eckpunktepapier.pdf?__blob=publicationFile&v=3**

Calvo Tello, J., Funk, S., Göbel, M., Kurzawe, D., Rißler-Pipka, N., Veentjer, U., & Weimer, L. (2023). Between corpora, tools and authority files: TextGrid Repository for Hispanic Studies. ***Revista de Humanidades Digitales*****,** ***8*****, 90–108.** **https://doi.org/10.5944/rhd.vol.8.2023.37994**

Calvo Tello, J., Rißler-Pipka, N., Chambers, S., Genêt, P., & Helling, P. (2024a, May 16). ***Metadata Desiderata for Literary Corpora: A survey for starting a conversation between literary studies, libraries and research data repositories.*** **https://doi.org/10.5281/zenodo.11202132**

Calvo Tello, J., Rißler-Pipka, N., Weimer, L., Veentjer, U., Kurzawe, D., Klammer, R., Göbel, M., Funk, S., Dogaru, G., Buddenbohm, S., & Barth, F. (2024b, August 5). ***Responsibility through FAIRness or why you Should Bring your Texts to TextGrid Repository*****. DH2024 (CLS Workshop – 'How To Do Research Responsibly'), Arlington.**

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, *19*(1). https://doi.org/10.5334/dsj-2020-043

CoreTrustSeal. (n.d.). *Website CoreTrustSeal*. Retrieved 25 July 2025, from https://www.coretrustseal.org/

DataCite Metadata Working Group. (2024, December 5). *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs (Version 4.6)*. DataCite e.V. https://doi.org/10.14454/csba-e454

DataGEMS Consortium. (n.d.). *Website DataGEMS*. Retrieved 28 July 2025, from https://datagems.eu/

DBpedia Association. (2025, June 20). *Website DBpedia*. https://www.dbpedia.org/

DCMI Usage Board. (2020). *DCMI Metadata Terms*. DublinCore. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Deutsche Forschungsgemeinschaft. (2024). *Guidelines for Consortia National Research Data Infrastructure (Form nfdi100)*. https://www.dfg.de/resource/blob/168700/nfdi100-en.pdf

DH-Kolloquium – Berlin-Brandenburgische Akademie der Wissenschaften**. (n.d.). Retrieved 1 July 2025, from https://www.bbaw.de/bbaw-digital/dh-kolloquium**

DIN-Normenausschuss Terminologie (NAT). (n.d.). *NA 105-00-06 AA: Sprachressourcen und Sprachtechnologie*. https://www.din.de/de/mitwirken/normenausschuesse/nat/nationale-gremien/wdc-grem:din21:63074672

Dogaru, G. (2023, June 22). *Das Text+-Langzeitarchiv: Eine generische Lösung für den nachhaltigen Erhalt von Daten in den Geisteswissenschaften*. https://doi.org/10.5281/zenodo.8108792

Dönicke, T., Barth, F., Varachkina, H., & Sporleder, C. (2022). MONAPipe: Modes of Narration and Attribution Pipeline for German Computational Literary Studies and Language Analysis in spaCy. In R. Schaefer, X. Bai, M. Stede, & T. Zesch (Eds), *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)* (pp. 8–15). KONVENS 2022 Organisers. https://aclanthology.org/2022.konvens-1.2/

Dumont, S., & Fechner, M. (2014). Bridging the Gap: Greater Usability for TEI encoding. *Journal of the Text Encoding Initiative*, *Issue 8*. https://doi.org/10.4000/jtei.1242

Ďurčo, M., Barbot, L., Tasovac, T., Tóth-Czifra, E., Roi, A., Chambers, S., Benardou, A., Constantopoulos, P., Dobreva, M., Scharnhorst, A., Kálmán, T., Rißler-Pipka, N., & McConville, A. (2023, November 22). *DARIAH Services Policy*. https://doi.org/10.5281/zenodo.10184825

Eberl, F., Jansen, L., Miller, B., Amelung, L., Danabalan, R., Demandt, É., Deschler, K., Eggert, M., Engel, J., Espinoza, S., Ferenz, S., Fritzsche, F., Goedicke, M., Götz, B., Hennig, C., Hofmann, A., Hoffmann, C., Hunold, J., Idda, T., …(2025). *Collaborative work in NFDI*. https://doi.org/10.5281/zenodo.8296724

Eckart, T., Buddenbohm, S., Steckel, A., & Wieder, P. (2025). *Text+ Architecture v2.0 (2025)* [Graphic]. Zenodo. https://doi.org/10.5281/zenodo.16036448

Eckart, T., Herold, A., Körner, E., & Wiegand, F. (2023). A Federated Search and Retrieval Platform for Lexical Resources in Text+ and CLARIN. *Electronic Lexicography in the 21st Century (eLex 2023 Conference Proceedings)*, 280–292.

ediarum. Digitale Editionen erstellen und publizieren. (n.d.). Retrieved 3 July 2025, from https://www.ediarum.org/

***Edit empfiehlt – Text+ Blog*. (2025, May 8). https://textplus.hypotheses.org/tag/edit-empfiehlt**

EOSC Association. (n.d.). *European Open Science Cloud*. Retrieved 28 July 2025, from https://eosc.eu/

EOSC Association. (2025a). *EOSC Federation Handbook*. Zenodo. https://doi.org/10.5281/zenodo.14999577

EOSC Association. (2025b, August 3). EOSC's build up phase to begin with March kickoff. https://eosc.eu/news/eoscs-build-up-phase-to-begin-with-march-kick-off/

European Commission, D. G. for C. N., Content and Technology. (2024, January 11). *Data Act | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/policies/data-act

European Parliament, & Council of the European Union. (2024, June 13). *Regulation (EU) 2024/1689 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)* [Official Journal of the European Union]. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

ExplosionAI GmbH. (n.d.). *spaCy · Industrial-strength Natural Language Processing in Python*. Retrieved 29 July 2025, from https://spacy.io/

FE / Web Annotation Protocol Compliant Annotation Store · GitLab. (2025, April 22). GitLab. https://gitlab.gwdg.de/fe/wapcas

Feidicker, C., Kaiser, J., & Calvo Tello, J. (2024, February 8). Den Überblick behalten: Annotation der User Stories mit der Basisklassifikation in Zotero. *Text+ Blog*. https://doi.org/10.58079/vs8p

FID-LG. (2022, August 30). *Aktionsplan 2022–2024 des FID-Lenkungsgremiums für das FID-Netzwerk*. https://wikis.sub.uni-hamburg.de/webis/images/8/8a/Aktionsplan-FID-Netzwerk_2022-2024.pdf

Fischer, B. K., Buddenbohm, S., & Annisius, M. (2023, May 19). *Die Gemeinsame Normdatei (GND)–The Integrated Authority File in Text+ as semantic link to DARIAH-EU*. https://doi.org/10.5281/zenodo.7949899

Fišer, D., & Witt, A. (Eds). (2022). *CLARIN: The infrastructure for language resources* (Vol. 1). De Gruyter.

Föhr, P., Buddenbohm, S., & Steckel, A. (2023). *Text+ Werkschau* [Graphic]. https://doi.org/10.5281/ZENODO.13987203

Franzini, G. (2022). *Catalogue Digital Editions* [Dataset]. https://dig-ed-cat.acdh.oeaw.ac.at/

Friedrichs, S., & Jander, M. (2022, December 9). *Der Text+-Helpdesk*. Text+ Plenary 2022, Mannheim. https://doi.org/10.5281/zenodo.7252932

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Geiger, J., Steiner, P., Desouki, A. A., & Lange, F. (2024). *DALIA Interchange Format* (Version 1.3). Zenodo. https://doi.org/10.5281/ZENODO.11521029

Genêt, P., Gradl, T., Hensen, K., Kudella, C., & Schulz, D. (2023a, October 5). *F wie Registry. Die Text+ Registry als Hilfsmittel zur Auffindbarkeit von Ressourcen*. FORGE 2023, Tübingen.

Genêt, P., Hensen, K., Herold, A., & Trippel, T. (2025, July 28). *Beispiele für Text+ Use Cases*. https://doi.org/10.5281/zenodo.16529536

Genêt, P., Hug, M., Leinen, P., & Trippel, T. (2023b, September 30). *Konzept und prototypische Implementierung beispielhafter Ressourcen*. Zenodo. https://doi.org/10.5281/zenodo.12771255

German National Library (DNB). (n.d.a). *GND - Gemeinsame Normdatei*. Retrieved 28 July 2025, from https://gnd.network/Webs/gnd/DE/Home/home_node.html

German National Library (DNB). (n.d.b). *Nestor Seal for Trustworthy Digital Archives*. Retrieved 29 July 2025, from https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/siegel.ht

GKFI. (n.d.). *Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V.* Retrieved 1 August 2025, from https://www.forschungsinfrastrukturen.de/doku.php/start

Glombiewski, N., Boukhers, Z., Beilschmidt, C., Drönner, J., Mattig, M., Piet, A., Pietrzynski, R., Jaberansary, M., Maia, M., Beyvers, S., Yediel, Y. Ü., Akhtar, M. H., Oberkampf, H., Hartman, J., Seeger, B., & Lange, C. (2024). *From Theory to Practice: Demonstrators of FAIR Data Spaces Across Different Sectors*. arXiv. https://doi.org/10.48550/ARXIV.2412.04969

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

Hedges, M., Neuroth, H., Smith, K. M., Blanke, T., Romary, L., Küster, M., & Illingworth, M. (2013). TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship. *Journal of the Text Encoding Initiative*, *5*, Article Issue 5. https://doi.org/10.4000/jtei.774

Henriksson, E., Myntti, A., Hellström, S., Erten-Johansson, S., Eskelinen, A., Repo, L., & Laippala, V. (2024). From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations. In M. Hämäläinen, E. Öhman, S. Miyagawa, K. Alnajjar, & Y. Bizzoni (Eds), *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* (pp. 308–318). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.nlp4dh-1.30

Henzen, C., Neidiger, C., Söding, E., & Trippel, T. (2025). *Report on the first NFDI Metadata Workshop*. Zenodo. https://doi.org/10.5281/zenodo.15227236

HERMES. (n.d.). Retrieved 1 July 2025, from https://hermes-hub.de/

Holmes, M., & Takeda, J. (2023). From Tamagotchis to Pet Rocks: On Learning to Love Simplicity through the Endings Principles. *Digital Humanities Quarterly*, *17*(1).

Hinrichs, E., Geyken, A., Leinen, P., Speer, A., Stein, R., Blumtritt, J., Borek, L., Eckart, T., Engelberg, S., Grötschel, M., Henrich, A., Heyer, G., Horstmann, W., Jefferies, N., Kudella, C., Lobin, H., Müller-Spitzer, C., Neuber, F., Neuefeind, C., … Witt, A. (2022). *Text+: Language- and text-based Research Data Infrastructure*. https://doi.org/10.5281/zenodo.6451706

Hinrichs, E., Hinrichs, M., & Zastrow, T. (2010). WebLicht: Web-Based LRT Services for German. In S. Kübler (Ed.), *Proceedings of the ACL 2010 System Demonstrations* (pp. 25–29). Association for Computational Linguistics. https://aclanthology.org/P10-4005/

***I-D-E – Institut für Dokumentologie und Editorik*. (2025, June 27). https://www.i-d-e.de/**

IAM4NFDI Project Consortium. (2024, November 21). *NFDI AAI documentation: Community AAI and infrastructure architecture*. IAM4NFDI (Base4NFDI). https://iam.services.base4nfdi.de/

ICOM-CIDOC LIDO Working Group. (2021, December 20). *LIDO - Lightweight Information Describing Objects Version 1.1*. https://www.lido-schema.org/schema/v1.1/lido-v1.1.html

Kett, J., Kudella, C., Rapp, A., Stein, R., & Trippel, T. (2022). Text+ und die GND – Community-Hub und Wissensgraph. *Zeitschrift Für Bibliothekswesen Und Bibliographie*, *69*(1–2), 37–47. https://doi.org/10.3196/1864295020691262

Klein, W., & Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). *Lexicographica*, *26*(2010), 79–96. https://doi.org/10.1515/9783110223231.1.79

Konsortialversammlung of the German National Research Data Infrastructure association. (2022, February 15). *Stellungnahme der NFDI-Konsortien zu Basisdiensten*. Zenodo. https://doi.org/10.5281/zenodo.6091657

Körner, E., Eckart, T., Helfer, F., & Kretschmer, U. (2024). Federated Content Search: Advancing the Common Search Infrastructure. In V. Vandeghinste & T. Kontino (Eds), *CLARIN Annual Conference Proceedings* (pp. 41–45). https://www.clarin.eu/sites/default/files/CLARIN2024_ConferenceProceedings_final.pdf

Körner, E., & Eckart, T. (forthcoming). Accessing linguistic content in distributed research environments. In P. Bánski, U. Heid, & L. Herzberg (Eds), *Harmonizing language data: Standards for linguistic resources*. DeGryuter.

Körner, E., Eckart, T., Herold, A., Wiegand, F., Michaelis, F., Bremm, M., Cotgrove, L., Trippel, T., & Rau, F. (2023, May). *Federated Content Search for Lexical Resources (LexFCS): Specification*. Zenodo. https://doi.org/10.5281/zenodo.7986303

Lemke, K., König, S., Lordick, H., & Brüning, G. (2025, February 10). *Empfehlung zur Erstellung, Bearbeitung und Publikation FAIRer Forschungsdaten in der Datendomäne Editionen – A Living Handbook über das How-to der Bewertung und Gewährleistung FAIRer, qualitativ hochwertiger editionswissenschaftlichen Forschungsdaten*. GitLab. https://gitlab.gwdg.de/textplus/textplus-editions/guidelines_sde

Lordick, H. (n.d.). *GND BEACON Hub*. GND BEACON Hub. Retrieved 30 June 2025, from https://dh-fx.de/gnd-beacon-hub

Lordick, H., & Hegel, P. (2024). BEACON. In S. Galka & H. W. Klug (Eds.), *KONDE Weißbuch*. https://www.digitale-edition.at/o:konde.254

Marčić, A., Hensen, K., Büdenbender, S., Lordick, H., Neuefeind, C., & Sievers, M. (2023, September 4). (Not) another Registry. Release Notice: Data Model for Curated Software Platform Version 1.0 [Billet]. *Text+ Blog*. https://doi.org/10.58079/uqh5

Mathiak, B., Heyer, G., Widmann, H., Henzen, C., & Czerniak, A. (2024). Search and Harvesting across NFDI Consortia – Gaps and Challenges. *Datenbank-Spektrum*, *24*(2), 77–84. https://doi.org/10.1007/s13222-024-00480-7

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). *Large Language Models: A Survey* (No. arXiv:2402.06196). arXiv. https://doi.org/10.48550/arXiv.2402.06196

NFDI e.V. (n.d.a). *Community Meetings*. https://www.nfdi.de/community-meetings/?lang=en

NFDI e.V. (n.d.b). *Humanities@NFDI*. https://www.nfdi.de/humanitiesnfdi/?lang=en

NFDI e.V. (n.d.c). *Sections*. https://www.nfdi.de/sections/?lang=en

NFDI e.V. (n.d.d). *Vision & Mission*. https://www.nfdi.de/association/?lang=en

NFDI e.V. (Section Ethical, Legal & Social Aspects). (2022, May). *Opinion on the EU Data Act Proposal* [Text]. European Commission - Have Your Say. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13045-Data-Act-amended-rules-on-the-legal-protection-of-databases/F3258672_en

NFDI e.V. (Ethical, Legal & Social Aspects). (2023a, May). *Stellungnahme zum Daten-Governance-Gesetz*. https://www.nfdi.de/wp-content/uploads/2023/05/ELSA-Stellungnahme-DGG.pdf

NFDI e.V. (Ethical, Legal & Social Aspects). (2023b, May). *Stellungnahme zur öffentlichen Konsultation zum Forschungsdatengesetz*. https://www.nfdi.de/wp-content/uploads/2023/05/NFDI-Stellungnahme-zum-Forschungsdatengesetz.pdf

NFDI4Culture Consortium. (n.d.). *Culture Knowledge Graph*. Retrieved 29 July 2025, from https://nfdi4culture.de/de/dienste/details/culture-knowledge-graph.html

NFDI4Objects. (2025, July 1). *DANTE – Datendrehscheibe für Normdaten und Terminologien*. NFDI4Objects. https://nfdi4objects.net//portal/services/dante/

NFDI Expert Committee. (2023). *Key points for the second funding period of the NFDI consortia*. Retrieved 2 August 2025, from https://www.dfg.de/resource/blob/204398/dfab28aa8bb3f2864cec55fb49370e93/eckpunkte-zweite-foerderphase-en-data.pdf

***nfdi.software*. (n.d.). Nfdi.Software. Retrieved 30 June 2025, from https://nfdi.software/software**

Neuber, F., Wiegand, F., & Hug, M. (2022a). *Multiple Dateninstanzen als Herausforderung für das Forschungsdatenmanagement – die Edition der Briefe von Jean Paul*. https://doi.org/10.5281/ZENODO.7377736

Neuber, F., Wiegand, F., & Hug, M. (2022b, December 6). »Eine einfältigere Frage an dich gab es wol nie als die, deren Antwort ich für mein Buch nöthig habe: Nämlich wie geht das Kinderlied weiter: Nikolaus, fang' die Maus …«. Task-Area-übergreifende Zusammenarbeit in Text+ am Beispiel der Forschungsdaten der Jean Paul Briefedition [Text/html]. *Text+ Blog*. https://doi.org/10.58079/uqg5

Open Archives Initiative. (n.d.). *Website Open Archives Initiative Protocol for Metadata Harvesting*. Retrieved 29 July 2025, from https://www.openarchives.org/pmh/

PhilFinder. (n.d.). PhilFinder. Retrieved 30 June 2025, from https://philportal.de/philfinder

Pichler, A., Pagel, J., & Reiter, N. (2025). Evaluating LLM-Prompting for Sequence Labeling Tasks in Computational Literary Studies. In A. Kazantseva, S. Szpakowicz, S. Degaetano-Ortlieb, Y. Bizzoni, & J. Pagel (Eds), *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)* (pp. 32–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.latechclfl-1.5

Politze, M., Wieder, P., Schimmler, S., & Diepenbroek, M. (2025). *Concept for Setting up the Overall Architecture Working Group in the NFDI Section 'Common Infrastructures'*. https://doi.org/10.5281/zenodo.15210894

Rat für Informationsinfrastrukturen. (2025). *Leistung in Verantwortung: Zur Zukunft der wissenschaftlichen Informationsinfrastrukturen in Deutschland*. https://rfii.de/?p=12040

re3data.org. (2013). *Registry of Research Data Repositories*. https://doi.org/10.17616/R3D

Research Software Directory. (n.d.). Research Software Directory. Retrieved 30 June 2025, from https://research-software.cceh.uni-koeln.de/

REWISO+ / FID-Netzwerk für Rechts-, Wirtschafts- und Sozialwissenschaften – Webis. (n.d.). Retrieved 30 June 2025, from https://wikis.sub.uni-hamburg.de/webis/index.php/FID-Netzwerk_f%C3%BCr_Rechts-,_Wirtschafts-_und_Sozialwissenschaften_/_REWISO%2B

***RIDE – A Review Journal for Scholarly Digital Editions and Resources*. (n.d.). Retrieved 1 July 2025, from https://ride.i-d-e.de/**

Rißler-Pipka, N., & Weimer, L. (2023, May 23). *The Association for Research Infrastructures in the Humanities and Cultural Studies – an interface between national and international research infrastructures*. https://doi.org/10.5281/zenodo.7960885

Ritter, X. (2025). *KGI4NFDI*. Base4NFDI. https://base4nfdi.de/projects/kgi4nfdi

Sahle, P. (2020). *Digital Scholarly Editions* (Version 3.0) [Dataset]. https://v3.digitale-edition.de/

Schema.org. (n.d.). *Website Schema.org*. Retrieved 29 July 2025, from https://schema.org/

Schmidt, J., Herrgen, J., Kehrein, R., Lameli, A., & Fischer, H. (n.d.). *Regionalsprache.de (REDE III). Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Retrieved 29 July 2025, from https://www.regionalsprache.de/

Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann, M., & Röpke, J. (2016). *Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen*. https://doi.org/10.17175/2020_006

Schulz, D., Gradl, T., Eckart, T., & Buddenbohm, S. (2025, February 10). *Text+ Registry: Data model Editions, Approach of Text+ towards Ontology Harmonization: Presentation for the Working Group Ontology Harmonization 10.02.25*. https://doi.org/10.5281/zenodo.14833110

Sherpa, L., Attar-Khorasani, S., Sankar, R., Müller-Pfefferkorn, R., & Ghiasvand, S. (2024). KEENsight: Cloud Based Collaborative Environment for Streamlining Machine Learning Development. In *15th International Conference on Information,*

Intelligence, Systems & Applications (IISA) **(pp. 1-8), doi: 10.1109/IISA62523.2024.10786688**

Sikora, U. (2022). *entityXML* [Data set]. GitLab. https://gitlab.gwdg.de/entities/entityxml

SPARQL Endpoint—NFDI4Culture. (n.d.). Retrieved 1 July 2025, from https://nfdi4culture.de/resources/knowledge-graph.html

SSH Open Marketplace. (2025, July 29). *Resources tagged as Text+* [Catalogue]. SSHOMP. https://marketplace.sshopencloud.eu/search?order=score&q=textplus&f.keyword=Textplus

SSHOC Consortium. (n.d.). *Social Sciences & Humanities Open Cloud (SSHOC)*. Https://Sshopencloud.Eu/. Retrieved 29 July 2025, from https://sshopencloud.eu/

Stein, R. (2018). Wissenschaftsgeleitete Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften: Bericht zur Workshop-Reihe. *Konferenzband EVA Berlin 2018*, *25*, 238–240. https://doi.org/10.11588/arthistoricum.442

Stephan, R., Labahn, K., & Glasow, M. (2012). Vernetzung biographischer Online-Ressourcen – Catalogus Professorum Rostochiensium und Rostocker Matrikelportal. In B. Mittermaier (Ed.), *Mitglied der Helmholtz-Gemeinschaft Vernetztes Wissen – Daten, Menschen, Systeme* (pp. 259–270).

Synthesia Ltd. (n.d.). *Synthesia KI Video Generator*. Retrieved 29 July 2025, from https://www.synthesia.io/de

Text+ Consortium. (n.d.a). *GND Agency Text+–Text+* [National Research Data Infrastructure]. Retrieved 29 July 2025, from https://text-plus.org/en/themen-dokumentation/standardisierung/gnd-agentur/

Text+ Consortium. (n.d.b). *Text+ – National Research Data Infrastructure* [National Research Data Infrastructure]. Retrieved 29 July 2025, from https://text-plus.org/en/

Text+ Consortium. (n.d.c). *Text+ Registry* [National Research Data Infrastructure]. Retrieved 29 July 2025, from https://registry.text-plus.org/default/landing

Text+ Consortium. (n.d.d). *User Stories–Text+* [National Research Data Infrastructure]. Retrieved 29 July 2025, from https://text-plus.org/en/themen-dokumentation/user-storys-2020/

Text+ Consortium. (n.d.e). *Federated Content Search* [National Research Data Infrastructure]. Retrieved 29 July 2025, from https://fcs.text-plus.org/

Text+ Consortium. (n.d.f). *Fachinformationsdienste* [National Research Data Infrastructure]. Retrieved 29 July 2025, from https://text-plus.org/vernetzung/fid/

Text+ Consortium. (n.d.g). *Text+ Helpdesk* [National Research Data Infrastructure]. Retrieved 29 July 2025, from https://text-plus.org/helpdesk/

Text+ Consortium. (2024a). *Text+ Consortium Progress Reports National Research Data Infrastructure (NFDI)* (Progress Report No. B-1 Progress Report Part 1). https://www.dfg.de/resource/blob/344884/text-.pdf

Text+ Consortium. (2024b, November 15). *Data–Text+* [National Research Data Infrastructure]. https://text-plus.org/en/daten-dienste/daten/

Text+ Consortium. (2025a, January 23). *LLM-Training under TDM exceptions–Text+* [National Research Data Infrastructure]. https://text-plus.org/en/themen-dokumentation/legal/llm-training/

Text+ Consortium. (2025b, March 6). *Architecture–Text+* [National Research Data Infrastructure]. https://text-plus.org/en/ueber-uns/architektur/

Text+ Consortium. (2025c, March 6). *Text+ Centers* [National Research Data Infrastructure]. https://text-plus.org/en/ueber-uns/textpluszentren/

Text+ Consortium. (2025d, June 13). *Data Depositing–Text+* [National Research Data Infrastructure]. https://text-plus.org/en/daten-dienste/depositing/

Text+ Consortium. (2025e, July 6). *LLM Service–Text+* [National Research Data Infrastructure]. https://text-plus.org/en/daten-dienste/llm-service/

***Text+ Consulting Guidelines*. (2025, March 20). GitLab. https://gitlab.gwdg.de/textplus/textplus-consulting-guidelines**

Text Encoding Initiative. (2025). *The TEI Guidelines*. TEI: Guidelines for Electronic Text Encoding and Interchange. https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

The Endings Project: The Endings Project | Le projet Endings. (2025). https://endings.uvic.ca

The European Parliament and The Council Of The European Union. (n.d.). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. http://data.europa.eu/eli/reg/2016/679/oj

Text+ Task Area Editions. (2024). *Die Editionsregistry (ERY) und das zugrundeliegende Datenmodell – Dokumentation und Spezifikation*. https://doi.org/10.5281/zenodo.13379995

Triantaphyllou, E., & Polatidis, N. (2023). *ChatGPT and Similar AI Tools: Their Potential and Challenges in Today's World: Vol. Special issue*. Elsevier. https://www.sciencedirect.com/special-issue/100PNDB4FZV

Veentjer, U., Buddenbohm, S., Calvo Tello, J., Funk, S. E., Klammer, R., Rißler-Pipka, N., Steckel, A., Weimer, L., Dogaru, G., & Göbel, M. (2025). *Fluffy Publication*

***Workflow: Preserving Humanities Research Data with the TextGrid Repository*. Transformations: A DARIAH Journal. https://doi.org/10.5281/zenodo.14025198**

Voß, J., & Schindler, M. (2017). *Gbv.github.io/beaconspec/*. BEACON Link Dump Format. https://gbv.github.io/beaconspec/

W3C. (n.d.a). *OWL - Web Ontology Language*. W3 - Semantic Web Standards. Retrieved 29 July 2025, from http://www.w3.org/TR/owl-ref/

W3C. (n.d.b). *RDF - Resource Description Framework*. Retrieved 29 July 2025, from http://www.w3.org/TR/rdf11-schema/

W3C. (n.d.c). *SPARQL 1.1 Query Language*. Retrieved 29 July 2025, from https://www.w3.org/TR/2013/REC-sparql11-query-20130321/

Weimer, L., & Buddenbohm, S. (2023, March 7). re3data und Text+ oder: Wie finden meine Daten den Weg in das richtige Repositorium? *Text+ Blog*. https://doi.org/10.58079/uqgk

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., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wissenschaftsrat. (2025). *Strukturevaluation der Nationalen Forschungsdateninfrastruktur (NFDI)*. 108 pages. https://doi.org/10.57674/WCDC-6D36

ZfdG – Zeitschrift für digitale Geisteswissenschaften. (n.d.). Zeitschrift für digitale Geisteswissenschaften. Retrieved 1 July 2025, from https://zfdg.de/

Zinn, C., Hinrichs, M., & Hinrichs, E. (2022). Adapting GermaNet for the Semantic Web. *Proceedings of the 18th Conference on Natural Language Processing*, 41–47. https://aclanthology.org/2022.konvens-1.6.pdf

Appendix 4: Additional Comments on Data Sheet

The current data sheet builds on the initial version submitted as part of the 2024 interim report. The figures presented are cumulative since the beginning of the project funding. At the time of reporting, the definitions and scope of the indicators were not clearly established and, in many cases, remain vague or insufficiently defined. As a result, the technical processes needed to collect and evaluate these indicators—such as automated reporting mechanisms—were not part of the original project plan and are still under development. This led to challenges in areas such as estimating storage requirements for services and data, which cannot yet be determined comprehensively or automatically. Consequently, many entries in the data sheet are based on incomplete information provided by project partners and do not yet reflect a fully consolidated or validated data basis.

For this report, additional contextual information has been included where possible. These annotations aim to clarify how specific values were derived. However, due to space limitations, these explanations remain brief and cannot cover all methodological details.

The data collection process was supported by contributions from all project partners and includes:

- Activities and Dissemination: A compilation of activities, datasets, conferences, and dissemination efforts, collected by May, 31st, 2025.
- Project Bibliography: A collaboratively compiled bibliography of all publications, as of July 31st, 2025.
- Datasets Provided: Since the Text+ registry is not yet complete, partners were asked to report the total number of datasets they have provided.

On July 15th, 2025, a workshop was held with representatives from the NFDI office, NFDI governance, the NFDI Task Force Evaluation and Reporting, the federal and state governments, the DFG, and other stakeholders to further refine and stabilize the definitions of these indicators. The outcomes of this workshop are not yet reflected in the current data but are expected to significantly shape future reporting practices.

Additionally, the structural monitoring has been introduced in the work programme of this application in all task areas, coordinated in TA 6 AD-M3. Several indicators to be included in the data sheet are now linked to deliverables in this area, which will help improve the consistency and reliability of future data. They are also points of reference for measuring success as detailed in Section 2.2.

We expect that the data foundation will continue to improve in the coming years as definitions are clarified and technical infrastructure for data collection is further developed.

Below are brief detailed comments on the individual fields in the datasheet:

1001 and 1002: Publications were listed as closed if the project bibliography did not explicitly identify open access in the metadata field “rights”

1003 – 1006: Based on the information provided by the individual partners

1007 – 1009: Based on information provided by the individual partners. Variations from the previous report are due to data curation processes; criteria for identifying and counting duplicates have been revised.

1010: It is impossible to estimate the exact size of our target community. The figure provided is based on data from the Federal Statistical Office of Germany, including 194,727 students in philologic disciplines and 8,497 FTE researchers in the humanities at public institutions. Other estimation methods such as website visits, number of potential library users, etc. are conceivable.

1011: Text+ participates in all NFDI-wide task forces and all NFDI sections

1012: With established cooperations in joined groups of all consortia, the maximum number is already reached without bi- or multilateral cooperations.

1013: Established cooperations with FID and learned societies previously not included

1014 – 1016: Based on the information provided by the individual partners

1017: five new participants: University of Marburg, University of Applied Science Wismar, the University of Wuppertal, Freie Universität Berlin, Verbundzentrale (VZG) des Gemeinsamen Bibliotheksverbundes (GBV)

1018 – 1020: Counting of the milestone and deliverables

1021: documented via tickets in the Text+ Helpdesk

1022 and 1023: Based on the application and the survey of disciplines of the project related staff

2001 – 2003: corresponds to the total number of events held or participated in (indicators 1007 – 1009)

3001 and 3002: Based on the information provided by the individual partners and very heterogeneous due to varying counting methods and data granularity. The term “data set” is underspecified; at least every data set counted here has a PID

4001: Based on the information provided by the individual partners on a yearly basis. The definition of “1 user” currently depends on each partner’s interpretation and monitoring procedures. To ensure consistency and comparability, a harmonised definition and counting method for users should be developed in coordination with NFDI. Not every services monitors user numbers (same for 4002, and 5001 – 5109).

4002: Based on the information provided by the individual partners. Registered Users includes some services contributed by Text+ partners requiring registration. Some of these services are used by the interested public in and outside of Germany, though they cannot be distinguished here. To ensure consistency and comparability, a harmonised definition and counting method for registered users should be developed in coordination with NFDI and the Base Services IAM.

4003: Text+ uses and plans to use at least the Base4NFDI basic services

4004 – 4008: Based on the information provided by the individual partners

5001 – 5109: Based on the information provided by the individual partners/service provider. For various reasons not every services monitors user numbers. The monitoring of individual services is heterogeneous.

Appendix 5: Text+ supplementary data sheet for consortia of the National Research Data Infrastructure (NFDI)

Please note: The supplementary data sheet will adobe transmitted in the original formatting as XLS to the DFG. The supplementary data sheet below is part of the proposal submission packages.

Preface

1. Purpose and goal of the data sheet

This “Supplementary data sheet for consortia of the National Research Data Infrastructure (NFDI)” supplements the renewal proposals of the NFDI consortia with quantitative information. The data sheet is not intended to compare different consortia with one another. Its purpose is to provide a better understanding of the development of a consortium over the course of time. To this end, the data sheet is designed to help answer following important questions:

- What level of maturity has a consortium and the communities it addresses reached?
- What is the exact work, publication and service profile of a consortium and to what extent are the different services used?

Both the data sheet and the quantitative data it includes are to be considered neither detached from the associated renewal proposal nor, in and of themselves, as indicators of quality. Since all data reported in the data sheet are directly related to the respective renewal proposal, all kinds of background information, explanations, additions, etc. to the data should be noted in appendix 4 to the form nfdi110, not in the data sheet itself.

2. Instructions for completing the data sheet

Before filling out the data sheet, please read the document "Guide to filling out the Supplementary data sheet for consortia of the National Research Data Infrastructure (NFDI)" (www.dfg.de/nfdi_giude_data_sheet) carefully.

- In addition to the preface, the data sheet contains four further spreadsheets to be completed.
- The data sheet includes both free text fields, in which entries are to be made, and fields in which one of several given options must be selected per drop-down list.
- Some fields of the spreadsheet are linked to each other. For example, in some cases, the cell "ID" is automatically filled in after some other cells have been completed.

- Please do not change the pre-filled fields of the spreadsheet, the formatting or the structure of the spreadsheet.
- In order to prevent incorrect entries, the areas of the spreadsheet where no entries are to be made are locked.
- IDs marked with a yellow background colour do not have to be completed until the progress reports on the second funding phase of the consortium have to be submitted.
- Please do not list your own key performance indicators (KPI), additions, explanations, and other further details on the recorded data in the data sheet itself, but instead in appendix 4 to the form nfdi110.

3. Questions

Feel free to contact nfdi@dfg.de if you have any questions or uncertainties regarding the data sheet.

4. References

This data sheet refers to the following sources, which should to be acknowledged:

1. Amelung, L., Anthofer, V., Danabalan, R., Demandt, É., Ebert, B., Elschner, E., Espinoza, S., Eufinger, J., Fuchsloch, S., Götz, B., Henzen, C., Hunold, J., Idda, T., Jansen, L., Krieger, U., Rodrigues, C. M., Meister, M., Miller, B., Pitroff, S., … Zinke, W. (2023). White Paper: Interim Report Reference. Version 2. Zenodo. https://doi.org/10.5281/zenodo.10101412.
2. DataCite Metadata Working Group. (2021). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V. https://doi.org/10.14454/3w3z-sa82
3. FAIR Data Maturity Model Working Group. “FAIR Data Maturity Model. Specification and Guidelines”, June 25, 2020. https://doi.org/10.15497/rda00050.
4. “FAIRassist.org.” FAIRsharing team & University of Oxford. Accessed November 24, 2023. https://fairassist.org/.
5. German Research Foundation (DFG). “What is open access?” Accessed November 24, 2023. www.dfg.de/what_is_open_access
6. “KDSF – Standard für Forschungsinformation in Deutschland“. Version 1.3. Accessed November 24, 2023. https://kerndatensatz-forschung.de/index.php?id=speztables
7. Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In Journal of Web Semantics, 17: 33-43. https://doi.org/10.1016/j.websem.2012.08.001. Open Access at: http://speroni.web.cs.unibo.it/publications/peroni-2012-fabio-cito-ontologies.pdf.

8. Turewicz, Michael, Scholz, Uwe, Glöckner, Frank Oliver, Dammann-Kalinowski, Tanja, and Wittchen, Manuel. “De.nbi Service Category-specific KPI Selection and Criteria”. Zenodo, May 31, 2022. https://doi.org/10.5281/zenodo.6597826.
9. “Wikidata.org.” Accessed November 24, 2023. https://www.wikidata.org/.

Supplementary data sheet for consortia of the National Research Data Infrastructure (NFDI)		
0. General information		
ID	**Information**	**Entry**
0001	Name of the consortium	*Text+*
0002	Start of funding	10/1/2021
0003	DFG reference number	NFDI 25/1
0004	Applicant institution	Leibniz-Institut für Deutsche Sprache Mannheim
0005	Spokesperson	Prof. Dr. Andreas Witt
0006	Number of funding phase	1
0007	Submitted as a supplement to	Renewal Proposal
0008	Date of submission	5/31/2025
0009	Reporting period	10/21 – 05/25

Supplementary data sheet for consortia of the National Research Data Infrastructure (NFDI)		
Name of the consortium		***Text+***
1. Overarching information about a consortium		
ID	**Information**	**Entry**
1001	Number of scholarly works published in OPEN access	273
1002	Number of scholarly works published in CLOSED access	255
1003	Number of software items published in OPEN source	23
1004	Number of NON-open source software items published	0
1005	Number of educational media resources published in OPEN access	42
1006	Number of educational media resources published in CLOSED access	0
1007	Number of events HELD	26
1008	Number of events HELD FOR EDUCATIONAL PURPOSES	112
1009	Number of events PARTICIPATED IN	148
1010	Size of the target community in total	203224
1011	Number of participations in working GROUPS OR SECTIONS of the NFDI Association	14
1012	Number of cooperations with other NFDI CONSORTIA	26
1013	Number of cooperations with actors OUSTSIDE THE NFDI system	41
1014	Number of cooperations with actors OUTSIDE GERMANY	42

1015	Number of INSTITUTIONALISED cooperations	0
1016	Number of OTHER KINDS of cooperations	0
1017	Number of actors/communities/institutions newly integrated since the proposal was submitted	5
1018	Number of tasks COMPLETED	109
1019	Number of tasks FAILED	0
1020	Number of tasks that were ADAPTED OR MODIFIED during the funding period	17
1021	Number of user consultations	*758*

1022	**Communities/subjects/disciplines represented in the consortium**	**RB-Nr**	**Targeted communities/subjects/disciplines**	**RB-Nr**
1022100	*Art History*	1.13	*Greek and Latin Philology*	1.11
1022101	Musicology	1.13	Linguistics	1.14
1022102	General and Comparative Linguistics, Experimental Linguistics, Typology, Non-European Languages	1.14	Literary Studies	1.15
1022103	Individual Linguistics, Historical Linguistics	1.14	Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies	1.16
1022104	Applied Linguistics, Computational Linguistic	1.14	Philosophy	1.18
1022105	German Medieval Studies (Medieval German Literature)	1.15		
1022106	German Literary and Cultural Studies (Modern German Literature)	1.15		
1022107	European and American Literary and Cultural Studies	1.15		

1022108	General and Comparative Literary Studies; Cultural Studies	1.15		
1022109	History of Philosophy	1.18		
1022110	Theoretical Philosophy	1.18		
1022111	Operations Management and Computer Science for Business Administration	1.24		
1022112	Communication Technology and Networks, High-Frequency Technology and Photonic Systems, Signal Processing and Machine Learning for Information Technology	4.42		
1022113	Software Engineering and Programming Languages	4.43		
1022114	Data Management, Data-Intensive Systems, Computer Science Methods in Business Informatics	4.44		
1022115	Political Science	1.23		

Supplementary data sheet for consortia of the National Research Data Infrastructure (NFDI)

Name of the consortium	***Text+***

2. Information about a consortium's networking with the community

ID	Information	Entry
2001	Number of community-related activities	286

3. Overview of the research data of a consortium

ID	Information	Entry
3001	Number of datasets provided by the consortium	1097822
3002	Number of datasets fulfilling the quality criteria established by the consortium	1097822

4. Overview of the services of a consortium

ID	Information	Entry
4001	Number of users of all services provided by the consortium	17540700
4002	Number of registered users	477460
4003	Number of services reused that were developed by other consortia	*8*
4004	Number of OPEN SOURCE services	53
4005	Number of services in the development stage	2
4006	Number of services in the prototype stage	3
4007	Number of services in the operation stage	104
4008	Number of services terminated in the course of the funding phase	0

Information about each individual service of a consortium

Supplementary data sheet for consortia of the National Research Data Infrastructure (NFDI)				
Name of the consortium			***Text+***	
5. Information about each individual service of a consortium				
ID	**Name of the Service**	**Service Category**	**Key Performance Indicator**	**Entry**
5001	Docker Swarm Cluster	Storage	Total amount of provided space (in gigabytes)	x
5002	Elastic Cluster	Web Application	Total number of Visits	x
5003	Elastic Data Analysis Dashboards (Kibana)	Web Application	Total number of Visits	x
5004	etym.on	Web Application	Total number of Visits	x
5005	Formulae - Litterae - Chartae Werkstatt!	Web Application	Total number of Visits	x
5006	Forschungsdatenrepository	Storage	Total amount of provided space (in gigabytes)	x
5007	HPC Rechner (Nvidia GPU)	Storage	Total amount of provided space (in gigabytes)	x
5008	INEL Corpora	Web Application	Total number of Visits	x
5009	MeineDGS	Web Application	Total number of Visits	x
5010	OAI Provider	Web Application	Total number of Visits	x
5011	RAGestry	Web Application	Total number of Visits	x
5012	Archiving Workflow for Editions	Workflow	Total number of Executions	x
5013	Research Software Directory	Database	Total number of Visits	x
5014	Editions Registry	Database	Total number of Visits	x
5015	Text+ Research Rendezvous - Offene Sprechstunde von Text+	Web Application	Total number of Visits	x
5016	AGATE	Tool/Application	Total number of Downloads	x
5017	FCS MWB	Library/API	Total number of Downloads	x
5018	Text+ Helpdesk	Web Application	Total number of Visits	592
5019	DLA Data+	Library/API	Total number of Downloads	12600
5020	DLA Katalog Kallías	Data Curation	Total number of data sets curated	3900000
5021	DH-Projekte	Workflow	Total number of Executions	x
5022	DH-Stipendien	Workflow	Total number of Executions	x
5023	Digitale Sammlung	Storage	Total amount of provided space (in gigabytes)	750000
5024	DNB-Lab	Web Application	Total number of Visits	x
5025	GND	Database	Total number of Visits	x
5026	GND-Explorer	Tool/Application	Total number of Downloads	x
5027	Metadatendienste	Database	Total number of Visits	x
5028	DWA-App	Database	Total number of Visits	2700
5029	GOBA	Data Curation	Total number of data sets curated	2791600
5030	HNWb-App	Database	Total number of Visits	900
5031	Kartenkatalog	Database	Total number of Visits	x
5032	Lingurep	Data Curation	Total number of data sets curated	1300
5033	REDE SprachGIS	Database	Total number of Visits	437300
5034	REDE SprachGIS-Workshop	Training	Total number of Training activities	x

5035	SPARQL	Library/API	Total number of Downloads	x
5036	SprachGIS Helpdesk	Web Application	Total number of Visits	x
5037	Wenkerbogen-App	Data Curation	Total number of data sets curated	63746
5038	YouTube-Kanal	Training	Total number of Training activities	76
5039	Bagman	Web Application	Total number of Visits	x
5040	CMDI Explorer	Web Application	Total number of Visits	x
5041	GermaNet	Database	Total number of Visits	x
5042	GermaNet API	Library/API	Total number of Downloads	x
5043	GermaNet Crosswords	Web Application	Total number of Visits	x
5044	GermaNet FCS endpoint	Library/API	Total number of Downloads	x
5045	GermaNet-Rover	Web Application	Total number of Visits	138
5046	TALAR	Storage	Total amount of provided space (in gigabytes)	600
5047	TALAR OAI-PMH endpoint	Library/API	Total number of Downloads	x
5048	Tübinger Treebank Collection	Database	Total number of Visits	9009
5049	Tündra	Database	Total number of Visits	8467
5050	Tündra Treebanks FCS Endpoint	Library/API	Total number of Downloads	x
5051	WebLicht	Workflow	Total number of Executions	x
5052	WebLicht as a Service	Library/API	Total number of Downloads	33053
5053	WebLicht-Batch	Web Application	Total number of Visits	315999
5054	Repositorium der Herzog August Bibliothek	Data Curation	Total number of data sets curated	586
5055	Wolfenbütteler Digitale Bibliothek	Database	Total number of Visits	x
5056	edition2LD	Workflow	Total number of Executions	x
5057	Deutsches Referenzkorpus (DeReKo)	Database	Total number of Visits	x
5058	Endpunkt für FCS/Lex-Fcs	Web Application	Total number of Visits	x
5059	IDS-Repositorium	Storage	Total amount of provided space (in gigabytes)	6000
5060	KorAP	Web Application	Total number of Visits	x
5061	KorAP FCS-Endpunkt	Library/API	Total number of Downloads	x
5062	OWID FCS-Endpunkt	Library/API	Total number of Downloads	x
5063	Ressourcen-Import für Endpunkt	Tool/Application	Total number of Downloads	x
5064	Jupyter	Web Application	Total number of Visits	100000
5065	Hypotheses-DE	Web Application	Total number of Visits	13200000
5066	perspectivia.net	Web Application	Total number of Visits	1122029
5067	QED	Web Application	Total number of Visits	x
5068	Text+ Blog	Web Application	Total number of Visits	x
5069	Repository NAWG	Storage	Total amount of provided space (in gigabytes)	3,6
5070	CURL-Portal	Tool/Application	Total number of Downloads	9522000
5071	FCS Endpoint Validator	Tool/Application	Total number of Downloads	523200
5072	Repo SAW	Data Curation	Total number of data sets curated	33100
5073	Text+ Consulting/Support	Web Application	Total number of Visits	758
5074	FCS	Library/API	Total number of Downloads	3253200
5075	epidat (Inschriftenediteionen)	Storage	Total amount of provided space (in gigabytes)	400

5076	GND BEACON Hub	Storage	Total amount of provided space (in gigabytes)	20
5077	GND BEACON Overview	Web Application	Total number of Visits	x
5078	STI Linked Data Service	Web Application	Total number of Visits	x
5079	DARIAH-DE Publikator	Workflow	Total number of Executions	x
5080	DARIAH-DE Repository	Data Curation	Total number of data sets curated	3979
5081	Data Federation Architecture	Database	Total number of Visits	x
5082	Datasheet Editor	Tool/Application	Total number of Downloads	x
5083	DHd-Blog	Web Application	Total number of Visits	x
5084	Discuss Data	Web Application	Total number of Visits	x
5085	entityXML	Data Curation	Total number of data sets curated	x
5086	Geo-Browser	Web Application	Total number of Visits	x
5087	MONAPipe	Workflow	Total number of Executions	x
5088	TextGrid Repository	Data Curation	Total number of data sets curated	696596
5089	TG Import UI	Workflow	Total number of Executions	x
5090	tgadmin	Workflow	Total number of Executions	x
5091	tgclients	Workflow	Total number of Executions	x
5092	tg-model	Workflow	Total number of Executions	x
5093	Webportal	Web Application	Total number of Visits	10000
5094	Repositorium (Saarbrücken)	Storage	Total amount of provided space (in gigabytes)	80
5095	FDR	Data Curation	Total number of data sets curated	x
5096	HZSK Korpora	Database	Total number of Visits	x
5097	jOAI Provider	Library/API	Total number of Downloads	x
5098	Acoustic Database of Russian Regional Speech	Data Curation	Total number of data sets curated	100
5099	Cologne Digital Sanskrit Lexicons	Data Curation	Total number of data sets curated	3700
5100	Critical Pali Dictionery	Data Curation	Total number of data sets curated	100
5101	Database of Romani Dialects	Data Curation	Total number of data sets curated	100
5102	DCH Datenarchivierung	Storage	Total amount of provided space (in gigabytes)	52000
5103	DCH FDM Beratung	Workflow	Total number of Executions	x
5104	DCH Ressourcenbetreuung	Data Curation	Total number of data sets curated	500
5105	Kosh	Data Curation	Total number of data sets curated	30300
5106	Language Archive Cologne	Storage	Total amount of provided space (in gigabytes)	6000
5107	Language Documentation Consulting	Workflow	Total number of Executions	x
5108	Premodern Chinese mining terms	Data Curation	Total number of data sets curated	100
5109	VedaWeb	Data Curation	Total number of data sets curated	100