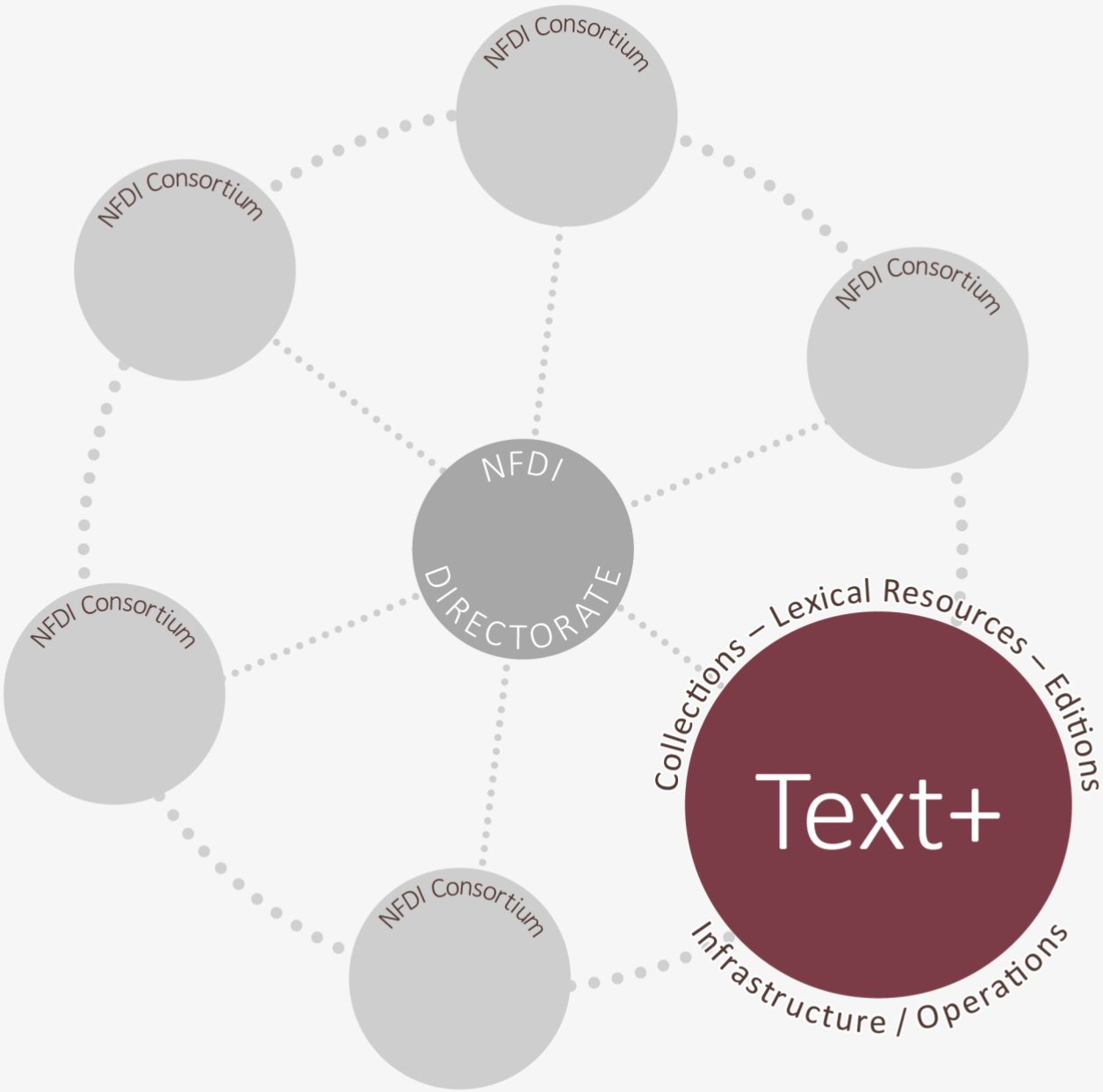

LANGUAGE- AND TEXT-BASED RESEARCH DATA INFRASTRUCTURE

Leibniz Institute for the German Language, Mannheim – Berlin-Brandenburg Academy of Sciences and Humanities, Berlin – German National Library, Leipzig and Frankfurt am Main – Göttingen State and University Library, Göttingen – North Rhine-Westphalian Academy of Sciences, Humanities and the Arts, Düsseldorf.

Academy of Sciences and Humanities in Hamburg – Academy of Sciences and Literature, Mainz – Bavarian Academy of Sciences and Humanities, Munich – German Literature Archive Marbach – German National Academy of Sciences Leopoldina, Halle (Saale) – Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen – Göttingen Academy of Sciences and Humanities – Heidelberg Academy of Sciences and Humanities – Herzog August Library Wolfenbüttel – Jülich Supercomputing Centre – Klassik Stiftung Weimar – Ludwig-Maximilians-University Munich – Max Weber Foundation, Bonn – Saarland University, Saarbrücken – Salomon Ludwig Steinheim Institute for German-Jewish History, Essen – Saxon Academy of Sciences and Humanities, Leipzig – Technical University of Darmstadt – University and State Library Darmstadt – Technical University of Dresden, Centre for Information Services and High Performance Computing – University of Applied Sciences Darmstadt – University of Bamberg – University of Cologne – University of Duisburg-Essen – University of Freiburg – University of Hamburg – University of Paderborn – University of Trier – University of Tübingen – University of Würzburg.

Table of Contents

List of Figures

List of Tables

List of Abbreviations[1]

AAI: Authentication and Authorisation Infrastructure

CARE: Collective benefit, Authority to control, Responsibility, and Ethics

CLARIN: Common Language Resources and Technology Infrastructure

DARIAH: Digital Research Infrastructure for the Arts and Humanities

DOI: Digital Object Identifier

ePIC: Persistent Identifiers for eResearch

FAIR: Findable, Accessible, Interoperable, Re-usable

FCS: Federated Content Search

FID: Specialised Information Service (*Fachinformationsdienst*)

GLAM: Galleries, Libraries, Archives, Museums

GND: Integrated Authority File (*Gemeinsame Normdatei*)

ITIL: Information Technology Infrastructure Library

NLP: Natural Language Processing

OCC: Infrastructure/Operations Coordination Committee

OCR: Optical Character Recognition

PID: Persistent Identifier

PM: Person Month(s)

RDM: Research Data Management

SCC: Scientific Coordination Committee

SDO: Standards Developing Organization

SFB: *Sonderforschungsbereich* (Collaborative Research Centre, CRC)

SRU: Search/Retrieve via URL

SSO: Single sign-on

TEI: Text Encoding Initiative

[1] Institutions are introduced directly in the text at their first appearance.

1 General Information

- Name of the consortium in English and German

  English: *Text+: Language- and text-based Research Data Infrastructure*

  German: *Text+: Sprach- und textbasierte Forschungsdateninfrastruktur*

- Summary of the proposal in English and German

English:

Text+ aims to develop a research data infrastructure for Humanities disciplines and beyond whose primary research focus is on language and text. Text+ will be flexible, scalable, and thus open for different discipline-specific requirements. By offering easy access to high-quality research data, Text+ will support a maximum of methodological diversity, which in turn is a prerequisite for innovative and transdisciplinary research.

Text+ focuses on Collections, Lexical Resources and Editions. These data domains have a long tradition of research and are linked to mature methodological paradigms that require distinctive but also cross-disciplinary practices of data generation, curation and management. The three types of research data are indispensable for a wide range of Humanities disciplines, including, but not limited to, Classical Philology, Linguistics, Literary Studies, Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies, Philosophy, and language- and text-based research in the Social and Political Sciences.

From the outset, 26 data centres will participate in Text+ that are technically sound and that are highly regarded in their fields of specialisation. They will provide data, tools, and services for the analysis and re-use of research data across a broad range of disciplines. By grouping data, tools, and services into thematic clusters, an optimal bundling is achieved.

There are 34 institutions participating in Text+ that represent the communities addressed by Text+ as broadly as possible: research libraries, universities, Digital Humanities data centres as well as members of the Union of German Academies of Arts and Sciences and of the Leibniz Society. In addition, leading computing centres ensure robust and persistent operation of services for a distributed research data infrastructure. The high level of interest in Text+ is not only evidenced by the substantial in-kind contributions by the Text+ partner institutions, but is also documented by the more than 120 research-driven user stories and by the large number of letters of support from the communities of interest participating in Text+.

At the heart of the governance structure are three scientific coordination committees for the data domains and one for the infrastructure. Their task is to continuously evaluate the portfolio of data,

tools and services and to promote its further development according to the priorities of the participating disciplines in coordination with the infrastructure providers.

The research data management strategy of Text+ is the core instrument for achieving the main objectives of Text+ in the NFDI context. It paves the way for the integration of data, tools and services into an infrastructure that meets relevant standards and implements the FAIR and CARE principles.

German:

Ziel von Text+ ist der Aufbau einer auf Text- und Sprachdaten ausgerichteten Forschungsdateninfrastruktur für die Geisteswissenschaften und für weitere sprach- und textbezogene Disziplinen. Text+ wird flexibel und skalierbar sein und damit offen für unterschiedliche disziplinenspezifische Anforderungen. Mit der Bereitstellung hochqualitativer Forschungsdaten wird Text+ ein Höchstmaß an methodologischer Vielfalt unterstützen, die wiederum Voraussetzung für innovative und transdisziplinäre Forschung ist.

Text+ legt den Fokus auf Sammlungen, lexikalische Ressourcen und Editionen. Diese Datendomänen haben eine lange Forschungstradition und sind mit ausgereiften methodologischen Paradigmen verknüpft, die charakteristische, aber auch bereichsübergreifende Praktiken der Erzeugung, Kuratierung und des Managements von Daten erfordern. Die drei Datendomänen sind unabdingbar für eine breite Palette von Fachdisziplinen, u. a. für die Klassische Philologie, die Sprach- und Literaturwissenschaft, Sozial- und Kulturanthropologie, Außereuropäische Kulturen, Judaistik und Religionswissenschaften, Philosophie sowie für die sprach- und textbasierte Forschung in den Sozial- und Politikwissenschaften.

Bereits zu Beginn von Text+ werden 26 fachlich ausgewiesene und technisch reife Datenzentren beteiligt sein. Sie stellen die Daten, Werkzeuge und Dienste für die Analyse und die Nachnutzung bereit und gewährleisten eine hohe fachliche Breite. Durch die Gruppierung von Daten, Werkzeugen und Diensten in thematische Cluster wird eine optimale Bündelung erreicht.

An Text+ beteiligen sich 34 Institutionen, die die von Text+ adressierten Fachdisziplinen in größtmöglicher Breite repräsentieren: Hochschulen, wissenschaftliche Bibliotheken, Datenzentren der Digital Humanities, Mitglieder der Deutschen Akademienunion und der Leibniz-Gemeinschaft. Dazu kommen führende Rechenzentren, die einen robusten und persistenten Betrieb der Dienste für eine distribuierte Forschungsdateninfrastruktur absichern. Das hohe Interesse an Text+ wird nicht nur durch die erheblichen Eigenmittel belegt, die die Institutionen zur Verfügung stellen, sondern wird auch durch über 120 forschungsgeleitete User Stories sowie die große Anzahl von Unterstützungsbriefen der beteiligten Fachdisziplinen für Text+ dokumentiert.

Im Zentrum der Lenkungsstruktur stehen drei wissenschaftliche Koordinationskomitees für die Datendomänen und eines für die Infrastruktur. Ihre Aufgabe ist, das Portfolio an Daten, Werkzeugen

und Diensten kontinuierlich zu evaluieren und seine Weiterentwicklung nach den Prioritäten der beteiligten Fachdisziplinen in Abstimmung mit den Infrastrukturanbietenden voranzutreiben.
Die Forschungsdatenmanagementstrategie stellt das entscheidende Instrument dar, um die übergeordneten Ziele von Text+ im NFDI Kontext umzusetzen. Sie ebnet den Weg für die Integration von Daten, Werkzeugen und Diensten in eine Infrastruktur, die relevanten Standards genügt und die FAIR und CARE Prinzipien umsetzt.

- Applicant institution

Applicant institution	Location	Short
Leibniz Institute for the German Language	Mannheim	IDS

- Spokesperson

Spokesperson	Institution, location
Prof. Dr. Erhard Hinrichs	Leibniz Institute for the German Language, Mannheim

- Co-applicant institutions

Co-applicant institutions	Location	Short
Berlin-Brandenburg Academy of Sciences and Humanities	Berlin	BBAW
German National Library	Leipzig and Frankfurt am Main	DNB
Göttingen State and University Library	Göttingen	SUB
North Rhine-Westphalian Academy of Sciences, Humanities and the Arts	Düsseldorf	NRWAW

- Co-spokespersons

Co-spokespersons	Institution, location	Task Area(s)
PD Dr. Alexander Geyken	Berlin-Brandenburg Academy of Sciences and Humanities, Berlin	Lexical Resources
Dr. Peter Leinen	German National Library, Leipzig and Frankfurt am Main	Collections
Prof. Dr. Andreas Speer	North Rhine-Westphalian Academy of Sciences, Humanities and the Arts, Düsseldorf	Editions

Regine Stein	Göttingen State and University Library, Göttingen	Infrastructure/Operations

-

- Participants

Participating institutions	Location	Short
Academy of Sciences and Humanities in Hamburg	Hamburg	AdWHH
Academy of Sciences and Literature, Mainz	Mainz	AdWMZ
Bavarian Academy of Sciences and Humanities	Munich	BAdW
German Literature Archive Marbach	Marbach	DLA
German National Academy of Sciences Leopoldina	Halle (Saale)	Leopoldina
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen	Göttingen	GWDG
Göttingen Academy of Sciences and Humanities	Göttingen	AdWGö
Heidelberg Academy of Sciences and Humanities	Heidelberg	AdWHD
Herzog August Library Wolfenbüttel	Wolfenbüttel	HAB
Jülich Supercomputing Centre	Jülich	JSC
Klassik Stiftung Weimar	Weimar	KSW
Ludwig-Maximilians-University Munich	Munich	LMU
Max Weber Foundation	Bonn	MWS
Saarland University	Saarbrücken	SLUni
Salomon Ludwig Steinheim Institute for German-Jewish History	Essen	STI
Saxon Academy of Sciences and Humanities	Leipzig	SAW
Technical University of Darmstadt (part of Darmstadt Cooperation, DACo)	Darmstadt	TUDa
Technical University of Darmstadt, University and State Library Darmstadt (part of Darmstadt Cooperation, DACo)	Darmstadt	USLDA
Technical University of Dresden, Centre for Information Services and High Performance Computing	Dresden	TUDD
University of Applied Sciences Darmstadt (part of Darmstadt Cooperation, DACo)	Darmstadt	h_da
University of Bamberg	Bamberg	UniBA
University of Cologne	Cologne	UniK
University of Duisburg-Essen	Duisburg	UniDUE
University of Freiburg	Freiburg	UniFR
University of Hamburg	Hamburg	UniHH
University of Paderborn	Paderborn	UniPB
University of Trier	Trier	UniTR
University of Tübingen	Tübingen	UniTÜ
University of Würzburg	Würzburg	UniWÜ

Contribution of the **AdWHH** in Task Area Collections: offering interdisciplinary and linguistically diverse language data collections including sign language in close co-operation with the Research Data Management department of the University of Hamburg.

Contribution of the **AdWMZ** in Task Area Editions: offering (1) a broad range of digital editions dealing with textual sources and materials from antiquity to the Avant-garde and (2) expertise in data modelling.

Contribution of the **BAdW** in Task Areas Collections, Lexical Resources, and Editions: (1) participating in the technical working groups of Text+, (2) providing digital or digitised dictionaries and (3) contributing metadata on its textual data.

Contribution of the **DLA** in Task Area Collections and Editions: contributing with collections of German electronic literature and digital-born material associated with German literature in the Science Data Center für Literatur (SDC4Lit).

Contribution of the **Leopoldina** in Task Area Editions: providing expertise in digital editions of textual data.

Contribution of the **GWDG** in all Task Areas, including Infrastructure/Operations: offering (1) computing and infrastructure with tailored IT solutions for Text+ and (2) cross-cutting technical services, including PID services, long-term preservation, AAI, cloud computing and storage.

Contribution of the **AdWGÖ** in Task Areas Lexical Resources and Editions: providing data and expertise on the historical lexicography of German.

Contribution of the **AdWHD** in Task Areas Collections, Lexical Resources, and Editions: offering (1) lexical data, (2) collections of inscriptions, (3) support for dissemination activities of Text+ and (4) participation in working groups.

Contribution of the **HAB** in Task Area Editions: offering (1) its own environment for editions linked to the local digital library, (2) training courses, (3) consulting services for external data providers, and (4) expertise on metadata for medieval manuscripts and old prints.

Contribution of the **JSC** in Task Area Infrastructure/Operations: offering (1) software for data analysis and data management, and (2) access to distributed storage and computer systems, especially the Text+ workspace concept.

Contribution of the **KSW** in Task Areas Collections and Editions: (1) participating in Text+ working groups, (2) contributing to the national and international dissemination of Text+, and (3) developing standards for textual data.

Contribution of the **LMU** in Task Area Collections: providing data and expertise, especially on spoken language and speech processing services.

Contribution of the **MWS** in Task Area Editions: (1) providing data and experience in creating editions, and community building, (2) contributing to the national and international dissemination of Text+ and user stories for Text+.

Contribution of **SLUni** in Task Area Collections: (1) providing basic consulting services and dissemination activities/teaching, (2) offering support for state-of-the-art language- and text-centred data analysis, and (3) hosting, e.g., multilingual and translation corpora.

Contribution of the **SAW** in Task Areas Lexical Resources and Operations/: (1) contributing digital material regarding lexical resources and (2) providing expertise on operations and technical infrastructure.

Contribution of **DACo** in Task Area Editions: providing expertise in the areas infrastructure development, teaching and training in the field of textual scholarship and digital editions.

Contribution of the **TUDD**, Centre for Information Services and High Performance Computing in Task Area Infrastructure/Operations: providing access to the High Performance Computing and Data Analytics infrastructure HRSK-II/HPC-DA.

Contribution of the **STI** in Task Area Editions: providing (1) data sets of numerous editions, prosopographic and bibliographic works, (2) authority files and controlled vocabularies for Linked Open Data, and (3) consulting on digital collections and editions.

Contribution of the **UniBA** in Task Area Infrastructure/Operations: offering (1) tools such as the Generic Search or the Data Modelling Environment and (2) expertise for interoperability, data and schema management.

Contribution of the **UniK** in Task Area Collections: providing (1) the repository for audio-visual data focusing on lesser-known and underdocumented languages, (2) consultation services and (3) adaptation of platforms and technologies to the specific needs of non-Western lexicography.

Contribution of the **UniDUE** in Task Area Collections: providing data and expertise for spoken language in political discourse.

Contribution of the **UniFR** in Task Area Collections: providing expertise on building and using resources for research on the Anglo-American and postcolonial English language and culture.

Contribution of the **UniHH** in (1) Task Area Collections: providing data and expertise for documenting language used in specific communicative situations, (2) in the Task Area Editions: providing experience in Sinology and Manuscriptology.

Contribution of the **UniPB** in Task Area Editions: offering teaching and training in the field of editorial processes.

Contribution of the **UniTR** in Task Area Lexical Resources: providing data and expertise, especially concerning technical challenges with regards to accessing lexical resources.

Contribution of the **UniTÜ** in Task Area Collections and Lexical Resources: (1) providing linguistically annotated data, (2) contributing lexical resources, expertise, technical services and (3) assisting in the administration and project management.
Contribution of the **UniWÜ** in Task Area Collections and Lexical Resources: supporting digitisation, corpus creation, digital editing, and digital analysis of texts.

Name	**Institution, location**	**Contribution of participant**
Prof. Dr. Andreas Henrich	University of Bamberg	Operations Vice Speaker
Prof. Dr. Verena Klemm	University of Leipzig	Scientific Coordination Committee Chair Editions
Prof. Vivien Petras, PhD	Humboldt University Berlin	Operations Coordination Committee Chair
Prof. Dr. Andrea Rapp	Technical University of Darmstadt	Scientific Vice Speaker
Prof. Dr. Sandra Richter	German Literature Archive Marbach	Scientific Coordination Committee Chair Collections
Prof. Dr. Ingrid Schröder	University of Hamburg	Scientific Coordination Committee Chair Lexical Resources
Prof. Dr. Elke Teich	Saarland University	Scientific Vice Speaker

Names and numbers of the DFG review boards (*DFG-Fachkollegien*) that reflect the subject orientation of the proposed consortium

101-02	Classical Philology
104	Linguistics
105	Literary Studies
106	Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies (partly)
108	Philosophy

2 Scope and Objective

2.1 Research Domains or Research Methods Addressed by the Consortium, Specific Aim(s)

Language and text are the basis of human culture, knowledge, and communication. As such, they underlie all aspects of society: everyday life, economy, education and research. Specific texts as well as languages are internationally protected cultural heritage, whose preservation and accompanying research is given a new dimension in the digital age. While "text as data" is a fairly recent phenomenon, it initiated a cultural change across the text- and language-based Humanities disciplines, ranging from Linguistics and Literary Studies to Cultural Anthropology and Philosophy. It is the overarching goal of Text+ to support and advance this change with a dedicated research data infrastructure and engage with other NFDI consortia that require text- and language-based resources and services as part of their research portfolios.

The primary research domains addressed by Text+ are those Humanities disciplines that create, access, collect, annotate, edit, analyse, visualise, publish, share, and archive text- and language-based research data in machine-readable form. Text+ will offer community-built solutions to manage and re-use research data for a wide range of languages. They include minority and endangered languages, multilingual resources, and resources for different text genres, such as scholarly publications, literary texts, letters, newspapers, inscriptions, public records, and non-fictional texts of various kinds. Text+ addresses different modalities, including written texts, speech, sign language, audio and video data, and derived text sources such as dictionaries and grammars.

The materials span several thousand years of cultural heritage, often collected over many years in academy-funded, long-term projects and with an ever-increasing amount of research data resulting from thousands of ongoing research projects in the Humanities. More text- and language-based research data are made available in very large quantities by research libraries and archives. Substantial parts of these materials are available in digital form. At the same time, many of the sources are not yet machine-readable or not yet digitised at all. The availability gap affects practically all areas of language and text data, but especially the small and endangered languages and literatures. Thus, high-quality digitisation, the curation of such data and standardisation of formats for digitised data remain a big challenge for any NFDI initiative in the Humanities.[2]

[2] It is commonly assumed that "growth in the volume and variety of data is mostly due to the accumulation of unstructured text data; in fact, up to 80% of all data is unstructured text data" (https://ec.europa.eu/info/sites/info/files/research_and_innovation/esfri-roadmap-2018.pdf, p. 108). This finding underscores the importance of including a consortium for language- and text-based research such as Text+ in a national research data infrastructure.

Initially, the Text+ consortium will focus on three types of research data, the Text+ data domains: language- and text-based Collections, Lexical Resources, and Editions. These three types of data have a long tradition in the Humanities and have given rise to mature methodological paradigms that require distinctive, yet cross-cutting practices of research data creation, curation and management. In addition, they also provide occasions for subject-specific and interdisciplinary methodological discussions and innovations. Thus, when we speak of "the three data domains", we do not refer to data resources or formats alone, but we refer to a common encoding of knowledge and information with a common set of research methodologies that cut across different Humanities disciplines and have significant relevance for a broad range of academic disciplines outside the Humanities. In essence, by shaping Text+ around established data domains rather than by reference to individual academic disciplines or to specific research methodologies, Text+ seeks to connect the diversity of research communities and facilitates answers to both disciplinary and interdisciplinary research questions.

In close collaboration with the communities of interest, Text+ has identified three Grand Challenges for building a research infrastructure that adheres to the FAIR[3] principles of Findable, Accessible, Interoperable, and Re-usable research data and the CARE[4] principles of Collective Benefit, Authority to Control, Responsibility, Ethics. Text+ is committed to addressing and mastering these Grand Challenges in the construction phase of the NFDI and beyond.

Grand Challenge I: Mastering the Diversity of Research Data and the Diversity of Communities of Interest for Language and Text

Research data in the Humanities are highly diverse in form and content. The data are contributed by a variety of stakeholders, including universities, non-university research institutes, research libraries, museums, archives, and commercial publishers. While large collections of textual data are typically housed in research libraries, archives, and non-university research institutes, research groups at universities contribute the long tail[5] of research data, which are often relatively small and distributed over many Humanities disciplines.[6] This long tail tends to be produced by individual researchers, by small, often interdisciplinary research teams, and by larger research initiatives with temporary funding. All these projects require institutional and external support for reliable research data management.

Any management strategy for language- and text-based research data in the Humanities must consider that Humanities research data are often housed in geographically distributed data repositories. Reasons for this distribution span from funding guidelines over institutional policies and regulations to

[3] https://www.force11.org/group/fairgroup/fairprinciples
[4] https://www.gida-global.org/care
[5] see Horstmann et al. (2017).
[6] The German Rectors' Conference (HRK) has identified the long tail of research data as one of the main challenges and opportunities for university repositories and long-term archiving. See HRK (2016).

copyright restrictions or other intellectual property constraints. Quite often, the data must remain at the host institution, which results in a distributed network of data repositories within Text+. It is therefore of primary importance to build a network of trust for research data providers and research data users alike to lay the foundation for a FAIR- and CARE-compliant Text+ infrastructure.

Such a network of trust must be based on a certification process with stringent criteria of (meta-)data quality, data archiving, and long-term availability of research data. This is essential for the certified data centres themselves but also for Humanities scholars who are seeking strong and recognised partners for developing and carrying out research data management plans, thereby meeting the expectations of funders and research communities to make their research data available.

The communities of interest in the Humanities come from many academic disciplines, including a high number of so-called *small disciplines* (*Kleine Fächer*), with different community practices of uptake of digital research data and associated research methodologies. Any NFDI consortium in the Humanities will need to respect this diversity of culture and will need to consider community-based strategies for research data management, community building, and community-based services.

The member institutions of Text+ jointly represent most, if not all relevant stakeholders that contribute research data in the Humanities. At the same time, the institutions participating in Text+ have closely collaborated with the communities of interest in order to identify their user requirements. The Text+ consortium is, therefore, in a strong position to formulate and carry out a research data management strategy (hereafter RDM strategy) that considers the diverse requirements of different types of data providers and data users.

Grand Challenge II: Assisting Humanities Scholars in the Re-use, Production, Preservation, and Innovative Use of Research Data

An NFDI consortium in the Humanities domain needs to be able to support the specific usage patterns for research data in its communities of interest.

(i) Data re-use: The most common type of user interaction that a research data infrastructure needs to support is easy access to existing data sources. Easy access entails the availability of a mature metadata infrastructure for finding research data and the adherence to domain-relevant community standards. In a distributed data infrastructure, access to the data sources themselves requires a federated search infrastructure with appropriate registries for data resources and associated software services. Consistent (cross-)referencing of metadata and research data presupposes access to authority files for data identification and data enrichment.

The Text+ institutions will make available a rich set of data resources, tools, and services for which authority files, metadata records and, in many cases, search portals for the data themselves have already been developed. This set of data resources and accompanying services will constitute the reference implementation of Text+. During the construction of the Text+ infrastructure, the main task

will be to integrate these resources into a common infrastructure and provide user-friendly federated access to metadata and object data. For this purpose, the infrastructure providers of Text+ can build on already developed relevant infrastructure components that can be re-used or further developed according to the changing needs of the communities of interest.

(ii) Data production: As described above, language- and text-based research data are produced in thousands of long tail research projects on a continuous basis. Making these data FAIR requires early and continuous interaction with data producers at all stages of the research data lifecycle. Such interactions need to be supported by the promotion and continuous refinement of standards and best practices for data encoding and by adherence to ethical and legal requirements. In addition, close user interaction requires a reliable data infrastructure as well as hands-on training events and consultation, the availability of published guidelines for data production, and the availability of software that can assist in the creation of data management plans.

The co-applicants and the participating institutions of Text+ have considerable experience in the promotion of and active participation in the development of standards and best practices for data encoding. Text+ will also be able to contribute published guidelines, training materials, and supporting data management software.

(iii) Data preservation: There are many legacy research data that are in danger of being irretrievably lost due to lack of funding, obsolete data formats, or poor data quality. Moreover, more and more funding agencies require data management plans and data preservation for a period of at least ten years after a research project has ended. An NDFI must offer data preservation services which avoid such data loss, and which give Humanities scholars the same stability that libraries and archives offer for printed volumes.

Text+ will offer data hosting and archiving services for external data whose integration into the Text+ infrastructure is sought by the creators of these research data or whose integration has been recommended by one of the communities of interest participating in Text+. The data repositories of the applicant and co-applicant institutions will play a crucial role in the implementation of these data hosting and archiving services of Text+.

(iv) Innovative use of research data: In order to be able to exploit the full potential of digital data resources for Humanities research, it is crucial to be open to new research approaches and ways of using the data. An infrastructure should offer suitable access options and interfaces as well as tools and services. Looking at the example of text and data mining (TDM), this could mean to provide access to texts in form of n-gram statistics, because this is the most meaningful download form possible due to legal or copyright issues. Using this access option, researchers can employ the data to develop new methods and approaches. At the same time the infrastructure should offer automatic text and data processing tools and services, allowing a broader community to use new methods. The availability of

such interfaces and services is essential for very large data resources housed in libraries and archives, where manual searches are simply not feasible due to the size of the data. At the same time, the ability to find patterns in such very large data sources is becoming increasingly important for data-driven research paradigms. In order to meet these challenges, automatic annotation tools for filtering the data, highly efficient algorithms for automatic querying, and automatic linking of tools and data will be essential.

Text+ will support Humanities scholars in innovative research projects by providing very large research data collections for written and spoken language with comprehensive access options and accompanying software tools. These will not only be of interest to Humanities scholars, but also to NFDI consortia in research domains that require automatic text processing services. At the same time, Text+ will communicate the risks of information bias inherent in research data to the users of Text+ materials, e.g. concerning the portrayal or the attitudes expressed toward political or social issues, toward ethnicity, or toward gender. It will develop techniques to identify and overcome such bias issues and will inform its communities of interest about such techniques.

Grand Challenge III: To Develop a Governance Model that Supports Joint Responsibility and Consensus Building among Infrastructure Providers and Infrastructure Users as well as Collaboration across Disciplinary Boundaries

The expert panel (*Expertengremium*) of the German Research Foundation (hereafter DFG) has identified six criteria as crucial for the success of the NFDI as a whole: (i) joint responsibility by all actors; (ii) early inclusion of scholars from the communities of interest; (iii) structural, organisational, and personal openness and dynamics; (iv) communication across disciplines; (v) thinking and acting in processes and structures rather than in terms of disciplines and project funding; (vi) co-operation instead of competition.[7]

In order to be able to meet these success criteria, Text+ has developed a governance model that is research-driven, inclusive, and allows easy integration of new types of research data and emerging research communities. At the core of this governance model, three Scientific Coordination Committees (hereafter SCCs) and one Infrastructure/Operations Coordination Committee (hereafter OCC) will continuously evaluate the Text+ portfolio of data, tools, and services. The SCCs will bring together experts in the areas of Collections, Lexical Resources and Editions, respectively. Membership in the SCCs will be determined in close consultation with the broad range of professional associations that have expressed their intent to actively participate in Text+. The SCCs and the OCC will monitor the initial portfolio offered by the Text+ institutions. At the same time, these committees will closely monitor the dynamics of digital research data production, shifts in research paradigms, as well as

[7] https://www.dfg.de/download/pdf/foerderung/programme/nfdi/stellungnahme_nfdi_eg.pdf

changes and innovations in research data management, thus enabling the continuous updating of the Text+ portfolio.

The RDM strategy of Text+ (section 4) and the Text+ work programme (section 5) centre around the three Grand Challenges outlined above and directly address the following six programme objectives for the NFDI as a whole: establishment of data handling standards, procedures and guidelines in close collaboration with the community of interest; development of cross-disciplinary metadata standards; development of reliable and interoperable data management measures and services tailored to the needs of the communities of interest; increased re-usability of existing data, also beyond subject boundaries; improved networking and collaboration with partners outside the German academic research system with expertise in research data management; involvement in developing and establishing generic, cross-consortia services and standards in research data management together with other consortia.[8]

2.2 Objectives and Measuring Success

Text+ offers a unique perspective on text and language encompassing a large range of different disciplines from Language Studies (Philology), Literary Studies and Linguistics to Anthropology, Cultural Studies and Philosophy. While the language- and text-oriented Humanities are characterised by a high diversity in disciplines, subjects, research foci and methods, in the last decades they have all been involved in a cultural change towards increasingly computationally supported analysis, ranging from the formal representation of contextual information by metadata to data-driven methods and probabilistic modelling. These methods differ, of course, across disciplines and/or subfields. For instance, in Dialect Studies relevant new methods come from machine learning (classification, clustering, aggregated distance measures) and shed new light on dialectal variation and underlying sociolinguistic factors.[9] In the field of Literary Studies, network analyses and scalable reading approaches open up new perspectives on literature.[10] For the first time, textual scholars can reconstruct and present entire networks of correspondences.[11] Even studies of the material artefacts in manuscript cultures have received an enormous boost through the digital accessibility of high-quality original images, standardised metadata and corresponding analysis methods.[12]

There is one major concern common to all these diverse efforts to adapt relevant computational methods and build suitable tools: high-quality, permanently citable data sets that are rich in relevant

[8] https://www.dfg.de/formulare/nfdi100/nfdi100_en.pdf, p. 4.
[9] see Nerbonne (2009).
[10] see Trilcke (2013) and Weitin/Werber (Eds.) (2017).
[11] see http://www.republicofletters.net/ and Hotson/Wallnig (Eds.) (2019).
[12] see, e.g., Horn (2020), Gülden et al. (2020), Tonne et al. (2019).

contextual information to allow for human interpretation.[13] This motivates the overarching goal of Text+:

Objective 1: Support Methodological Diversity by High-quality Research Data.

Text+ is the first large-scale multidisciplinary effort to share the rich and diverse expertise in the language- and text-oriented Humanities. The Text+ consortium is well prepared to embark on this endeavour.

First, we build on the combined results of two large infrastructure projects, CLARIN-D and DARIAH-DE, which include a base technical infrastructure (repositories, search), adherence to standards and community best practices (e.g. metadata) and a functioning, active academic network. This network manifests itself, inter alia, in the CLARIN and DARIAH Annual Conferences as well as in the wide spectrum of more than 120 authentic user stories.[14] These user stories were contributed by members of Text+'s communities of interest in response to a call issued by Text+. Here, the next steps are to reach out to communities that are not yet well represented and to make available more relevant resources and tools for data preparation, analysis and exploration including data visualisation.

The second pillar of expertise in Text+ are institutions dedicated to the (long-term) preservation of cultural heritage and to making research data for language and text available in digital form. Libraries and archives (or GLAM institutions) have made great efforts to make their collections digitally discoverable/findable and available through digitisation programmes and metadata indexing, thus helping to shape the digital turn in the Humanities. Here, the next steps are to close the still existing large availability gaps, to semantically index the resources through authority data and make them interoperable, and to convert image digitised material into full text digitised material (e.g. through OCR procedures) in order to be able to apply text and data mining, and semantic web technologies at a large scale. The academy programme of the Union of the German Academies of Sciences and Humanities currently comprises 137 long-term projects in which the global cultural heritage from thousands of years of human history is secured and researched in catalogues, registers, editions and dictionaries, both hybrid (print and digital), and purely digital. The academy community discusses and propagates standards and methods by regularly presenting results, e.g., at the annual conference of the working group e-Humanities.[15] Just as many representatives from GLAM institutions, they also participate in university teaching, in summer schools, and in working groups and conferences. These activities contribute to an intensive exchange between university and non-university institutions.

[13] This sets the language- and text-oriented Humanities apart from, e.g., Natural Language Processing, which primarily builds on big data that is typically not stratified by diasystemic linguistic variables.
[14] see https://www.text-plus.org/en/research-data/user-stories-en/
[15] see https://www.akademienunion.de/arbeitsgruppen/ehumanities/

The third pillar of expertise in Text+ is provided by its strong group of university participants. They represent a wide spectrum of disciplines, ranging from Computer Science, Literary Studies, (Computational) Linguistics, Medieval and Classical Studies, Philosophy, to a broad range of Language and Translation Studies. They have been pioneers in promoting digital scholarship in new university curricula and developing new research methods and paradigms in DFG Collaborative Research Centres (SFB 833, 980, 1102), DFG Priority Programmes (SPP 2207), and numerous individual research grants. In Text+, they make available their data and tools as a means for stimulating innovative research. Since university participants have a dual role as research data providers and as users of research data and services, they play an important role for interfacing with Text+'s communities of interest.

In sum, Text+ is in a unique position to stimulate and accompany the continued cultural change in the language- and text-oriented Humanities and to push for the next level of digital literacy in the communities represented as well as to transfer its knowledge and skills to interested, new communities. The field of Historical Linguistics, for instance, received a clear boost from the application of phylogenetic methods in recent years.[16] Similarly, distributional semantic methods[17] finally provided a computational approach to the Firthian notion of word meaning being determined by context[18] and opened up new perspectives on the construction and representation of lexical resources. Another more recent trend is to use mixed methods: corpus-based quantitative analysis is, for instance, combined with close-reading approaches (e.g. scalable reading[19]). While well-established in some fields of Linguistics, experimental techniques such as eye-tracking are slowly gaining traction in some parts of Literary Studies, e.g. in Reading Studies.[20] Again, the necessary condition to carry out these diverse endeavours and to embark on innovative research strands is the availability of high-quality data. Such data needs to be rich in information about relevant contextual variables of language use, such as register, genre, social group, gender, region, author or time. The data can be realised as an atlas of typological linguistic features and structures, a corpus based on clear data selection criteria, a critical edition or a lexical resource. Providing access to such data, extending the repertoire of relevant data, supporting interested communities in preparing high-quality data, and expanding their computationally supported analytic skills are the specific contributions of Text+. In this way, we expect to widen the potential for innovative and transdisciplinary research within the language- and text-oriented Humanities and beyond.

Due to the extensive prior work and collaboration of its partners and participants, Text+ will be in a strong position to address the three Grand Challenges identified in section 2.1 above. Text+ will

[16] see Gray et al. (2009).
[17] see, e.g., Mikolov et al. (2013).
[18] see Firth (1957).
[19] see Weitin/Werber (Eds.) (2017).
[20] see Wallot et al. (2013).

significantly expand and elaborate on previous offers in terms of text and language resources, technical services and community building. From the Grand Challenges, we derive the following further objectives which are in turn realised by Text+'s activities and measures as laid out in the work programme (see section 5).

Objective 2: Comply with Research Priorities of Communities of Interest

Research infrastructures all too often fall into the trap of pursuing a strategy: "If you build it, they will come."[21] In response to **Grand Challenge III**, Text+ wants to safeguard against such a supply-driven objective and wants to make sure that the selection of its research data and services is demand-driven and adheres to the research priorities of its communities of interest. The key to reaching this objective is a governance structure that will enable broad participation of all communities of interest and that will put in place a community-driven decision-making process. More specifically, the Text+ budget proposal contains flexible funds, which have not been assigned to specific institutions yet. They are earmarked for expansion of the Text+ data and service portfolio and will be allocated to additional participants on an as-needed basis. The needs and the priorities of such additional data and services will be assessed by three SCCs, one for each of the three data domains. These committees will consist of elected representatives of the professional associations that co-operate with Text+. The governance structure of Text+ is described in more detail in section 3.4. The SCCs will apply a set of strong criteria for data selection (see section 5 below) when they assess candidate data and services for inclusion in the Text+ portfolio.

Objective 3: Foster Transdisciplinary Co-operation

In response to **Grand Challenge I**, Text+ aims to integrate as many (Humanities) disciplines as possible that focus on language and text as their primary research data. Such openness appears crucial given the large number of Humanities disciplines and given the constraints of the NFDI, which strictly limits the number of consortia to a maximum of thirty. Fostering transdisciplinary co-operation is the key to success for such a large-scale integration of disciplines. At the same time, this co-operation will need to be sensitive to different levels of preparedness and discipline-specific requirements among individual communities of interest. Such requirements derive in part from the diversity of cultures, languages, writing systems, and written versus oral traditions. This diversity has been a focus in the many long-term research projects within the German Academies. The strong representation of German Academies in the Text+ consortium and the inclusion of distinguished individual participants in Text+ will ensure that the diversity of languages and cultures will be adequately represented.

[21] see van Zundert (2012).

Text+'s focus on collections, lexical resources, and editions will act as a nucleus for such co-operation, as will the large number of academic disciplines that are represented in Text+.

Objective 4: Advance Innovative Research

Text+ considers the development of a scalable research data infrastructure not as a goal in itself, but rather as a means toward advancing Humanities research as such and as a means of advancing innovative research paradigms and finding new answers to existing and future research questions (see **Grand Challenge II**). In order to facilitate and motivate innovative research, Text+ will solicit and advance user stories from its communities of interest on a regular basis and organise community events that target new directions for research methods and research questions. The range of topics and the large number of more than 120 authentic user stories contributed to Text+ in preparation of this funding proposal attest to the keen interest and pressing needs of Humanities scholars. They are accessible on the Text+ website for further review.[22] Throughout this proposal we will make reference to specific user stories that underscore the relevance of particular measures planned by Text+.

For an infrastructure provider, the user stories are a tremendous source of information on user needs and the quality and suitability of existing infrastructure components. Also, the user stories indicate directions for innovative research that can be enabled or facilitated by existing Text+ resources and services – or require new developments.

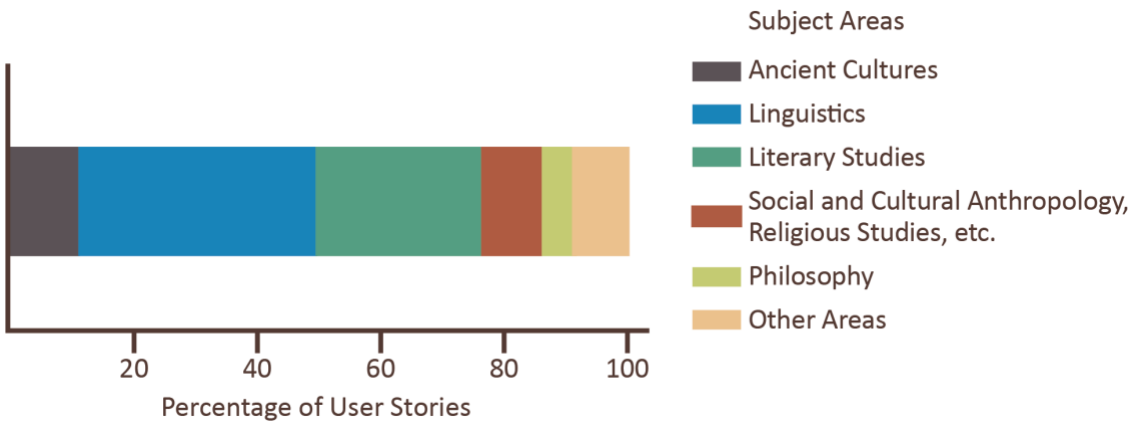

Figure 2.1 Percentage of user stories by subject area

Hence, the user stories will directly serve our overarching Objective 1 by providing us with fast feedback, including new trends in the diverse communities that Text+ addresses.

Apart from eliciting individual user stories, in the open call we issued in the summer of 2020, we also asked contributors to indicate their willingness to review the portfolio of Text+. In the planned regular calls for user stories, we will also ask for reviews, thus involving community members directly in

[22] https://www.text-plus.org/en/research-data/user-stories-en/

assessing components of the Text+ data and services portfolio. In this way, we gather basic information for more systematically designing and conducting user studies. Analyses according to different stakeholders and user types will provide information for the SCCs and OCC regarding the development of data and service portfolios.

Objective 5: Improve Research Data on a Massive Scale

Grand Challenge I underlines that research data in the Humanities suffer from considerable fragmentation and lack of coordination. In response to **Grand Challenges I** and **II**, Text+ will put in place a comprehensive strategy of data curation and data access that will address these shortcomings in compliance with the FAIR and CARE principles. The overall goal of this strategy is to improve findability and data access on a massive scale and to boost interoperability among data collections to facilitate data re-use. Text+ is well-positioned to carry out such a strategy because its member institutions have closely interacted with Humanities scholars and professional associations for many years. These interactions also included handling a variety of legal aspects (related to intellectual property, data protection and other issues) of text data access and re-use, providing legal information or manning legal helpdesks in relevant projects.

For each of the five objectives introduced above, a risk assessment and a strategy for risk management will be given in section 5.6. This strategy is based on the Measures defined in the different Task Areas of section 5 and is therefore better placed at the end of the proposal than in the current section.

3 Consortium

The following Text+ partners also participate in other consortia:

IDS: KonsortSWD; **BBAW**: NFDI4Objects; **DNB**: NFDI4Culture, NFDI4Objects, NFDI4Memory; **NRWAW**: NFDI4Culture, NFDI4Objects, NFDI4Memory; **SUB**: NFDI4Agri, NFDI4Biodiversity, NFDI4Culture, NFDI4Earth, NFDI4Memory, NFDI4Objects; **AdWGÖ**: NFDI4Memory; **AdWMZ**: NFDI4Culture, NFDI4Objects, NFDI4Memory; **BAdW**: NFDI4Objects, NFDI4Memory; **GWDG**: NFDI4Earth, NFDIxCS, NFDI4Biodiversity; **AdWHD**: GHGA, NFDI4Objects; **HAB**: NFDI4Memory; **H-DA**: NFDI4Culture, NFDI4Memory; **JSC**: DataPLANT, NFDI4Ing; **LMU**: DataPLANT, KonsortSWD, NFDI4Chem; **MWS**: NFDI4Culture, NFDI4Memory; **SLUni**: GHGA; **TUDa**: NFDI4Ing, NFDI4Objects, PUNCH4NFDI, NFDI4MobilTech, NFDIMatWerk, NFDIxCS; **TUDD**: Centre for Information Services and High Performance Computing: FAIRmat, GHGA, NFDI4Chem, NFDI4DataScience, NFDI4Earth, NFDI4Ing, NFDIxCS; **UniBA**: NFDI4Memory; **UniK**: NFDI4Culture, GHGA, NFDI4Chem; **UniDUE**: KonsortSWD; **UniFR**: DataPLANT; **UniPB**: NFDI4Culture; **UniTÜ**: DataPLANT, GHGA, NFDI4Earth; **UniWÜ**: NFDI4Chem.

3.1 Composition of the Consortium and its Embedding in the Community of Interest

As part of the preparations for the NFDI process, the applicant and co-applicant institutions of Text+ joined forces with a large number of professional associations in the Humanities and with other resource providers to engage in a broad and extended dialogue about the research data needs of Humanities scholars. These stakeholders participated in a series of three workshops entitled Research-driven Research Infrastructures for the Humanities and Cultural Studies in Germany (*Wissenschaftsgeleitete Forschungsinfrastrukturen für die Geistes- und Kulturwissenschaften in Deutschland*). The workshop series was co-organised by the infrastructure initiatives CLARIN-D and DARIAH-DE, by the Union of the German Academies of Sciences and Humanities, and by the professional association *Digital Humanities in the German-speaking World* (hereafter DHd). The needs of the communities of interest were articulated in individual position statements by the participating professional associations and by plenary discussions among all participants. The research data and service portfolios of the infrastructure providers were presented by a wide variety of institutions, including libraries, members of the Union of the Academies of Sciences, members of the Max Planck Society, members of the Leibniz Association, regional initiatives for research data management, Humanities data centres, and universities with a strong focus on Digital Humanities. The results of this workshop series are documented on the website of the workshop.[23] Building on the results of this workshop series, four different NFDI initiatives, NFDI4Culture, NFDI4Memory, NFDI4Objects, and Text+, were constituted. Each initiative focuses on different research data domains within the Humanities.

The composition of the Text+ consortium and the Task Areas align with the three data domains Collections, Lexical Resources and Editions and with the five objectives of Text+, which were introduced in section 2.2. The three data domains will be supported by a dedicated Task Area Infrastructure/Operations. These four Task Areas will be managed by the four co-applicant institutions of Text+ and their co-spokespersons. A fifth Task Area is devoted to Administration and will be managed by the applicant institution and its spokesperson. The applicant and co-applicant institutions of Text+ ensure the long-term viability of the infrastructure and contribute many years of experience in research data collection, management, and provisioning. In the Task Area Collections, the German National Library (hereafter DNB) in Leipzig and Frankfurt am Main will serve as co-applicant institution, with Peter Leinen as its co-spokesperson. The co-applicant institution Berlin-Brandenburg Academy of Sciences and Humanities (hereafter BBAW) and its co-spokesperson Alexander Geyken will coordinate the Task Area Lexical Resources. In the Task Area Editions, Andreas Speer, who represents the co-applicant institution North Rhine-Westphalian Academy of Sciences, Humanities and the Arts

[23] https://forschungsinfrastrukturen.de

(hereafter NRWAW), will serve as co-spokesperson. The co-applicant institution Göttingen State and University Library (hereafter SUB) will coordinate the Task Area Infrastructure/Operations, with Regine Stein acting as co-spokesperson, i.e. operations speaker. Spokesperson of the consortium, i.e. scientific speaker, will be Erhard Hinrichs for the Leibniz Institute for the German Language (hereafter IDS) in Mannheim, which acts as applicant institution for Text+ and will be responsible for the Task Area Administration. The Union of the German Academies of Sciences and Humanities has formed a working group on Digital Methods. This working group is coordinated by the BBAW and chaired by Martin Grötschel, the President of the BBAW. This will ensure an effective communication between Text+ and all member institutions of the Union of the German Academies.

The Text+ data domains are organised in **Text+ Thematic Clusters** (hereafter Clusters), which will provide comprehensive coverage of research data. The Clusters will bundle all activities related to specific subtypes of data and research methods in a data domain in accordance with the needs and research priorities of their specific communities of interest. They will engage in a continuous dialogue with Humanities scholars and offer data, software, and community services for a broad range of scientific disciplines in the Humanities whose research data focus on language and text. Initially, the following eight Clusters will focus on Ancient Cultures, Anthropology, Classics, Comparative Literary Studies, Computational Linguistics, Language and Literary Studies for European and non-European Philologies, Medieval Studies, Philosophy, and Religious Studies.

Figure 3.1 Data Domains and Initial Text+ Thematic Clusters

The Cluster *Contemporary Language*, the Cluster *Historical Texts*, and the Cluster *Unstructured Text* will offer a systematic approach to research data for Collections. For Lexical Resources, the Cluster *German Dictionaries in a European Context*, the Cluster *Born-Digital Lexical Resources*, and the Cluster *Non-Latin Scripts* will address a broad range of resources for contemporary and diachronic perspectives on language and text. For Editions, the Clusters *Ancient and Medieval Texts* and *Early Modern, Modern, and Contemporary Texts* will provide a layered service portfolio for a sustainable

research strategy due to the diversity of editorial models and their technical and methodological demands. All eight Thematic Clusters will provide a set of data services, community activities, and software services, which are described in detail in sections 4.4 and 5.

In the remainder of this section, the applicant and co-applicant institutions of Text+ are described in more detail with a particular emphasis on their contributions to the Text+ Clusters.

The **Leibniz Institute for the German Language (IDS)** in Mannheim, Germany, founded in 1964, is the leading national centre for research into and documentation of the German language in its contemporary usage and recent history. The mission of the IDS is to document, archive, and research the linguistic variety, structure, and use of the German language. Reference works (e.g. grammars and dictionaries) and computational language resources (especially large corpora[24] and analysis software) are both products of this research and created to support it. The IDS is also widely regarded as a hub of international German linguistics. Within CLARIN, the IDS hosts a certified data centre and focuses on standards for language resources and legal issues. Infrastructural and computational linguistic research activities[25] have recently been concentrated in the Department of Digital Linguistics.

Role: As applicant institution, the IDS will handle the Text+ budget and integrate its consortium of stakeholders. Taking the lead for the Task Area Administration, the IDS will be responsible for the disbursement of project funds to the co-applicant and participant institutions and will operate the Scientific Office of Text+. The IDS is also one of the central hubs in the Text+ Clusters, with two areas of specialisation.

For the data domain Collections, the IDS will coordinate the Cluster *Contemporary language in different modalities: written, spoken, and multi-modal* together with the University of Tübingen. First, the IDS will contribute its unique resources of contemporary German to the reference implementation, especially the (written) German Reference Corpus (DeReKo)[26], and corpora from the Archive for Spoken German. The IDS also provides interfaces to query and analyses the corpora: for spoken corpora, the Database for Spoken German[27] (12,000 registered users), and for written corpora (over 54,000 registered users) COSMAS II (developed since the 1990s) and KorAP[28] (Corpus Analysis Platform), which is optimised for large, multiple annotated corpora and complex search mechanisms and supports several query languages. Moreover, the IDS will provide legal expertise related to data protection as well as intellectual property and licensing issues.[29] It will also support standardisation activities building on its co-operation in numerous international committees, such as the Text Encoding

[24] see, e.g., Kupietz/Schmidt (Eds.) (2018) and the workshop series on *Challenges in the management of large corpora* (CMLC), starting with Bański et al. (Eds.) (2012).
[25] see, e.g., Lobin et al. (Eds.) (2018).
[26] see Kupietz et al. (2009, 2010, 2018b) and https://www1.ids-mannheim.de/kl/projekte/korpora
[27] see Schmidt et al. (2019) and http://dgd.ids-mannheim.de
[28] see Bański et al. (2013), Diewald/Margaretha (2016), Kupietz et al. (2017) and https://korap.ids-mannheim.de/
[29] see, e.g., Arnold et al. (2020), Kamocki/Witt (2020).

Initiative (hereafter TEI) and the International Standards Organization (hereafter ISO). The IDS will also contribute to requirements engineering and the integration of local endpoints and catalogues into the overall structure as well as to linking data. It brings in its experience in contributing significantly to, for instance, the CLARIN Virtual Collections Registry and the Federated Content Search.

The IDS will participate in all Measures of Collections.

For the data domain Lexical Resources, the IDS will contribute to the Cluster *German Dictionaries in a European Context*. The IDS will participate in Measure 1 and 4 of Lexical Resources.

The **Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)**, formerly Prussian Academy of Sciences, is an association of outstanding scientists and scholars from all over the world with over 300 years of tradition. It hosts about 25 world-wide renowned long-term projects, including the Dictionary of Ancient Egyptian,[30] the Digital Dictionary of the German Language (hereafter DWDS),[31] critical editions of the works of Gottfried Wilhelm Leibniz and Alexander and Wilhelm von Humboldt, the publications, manuscripts and correspondence of Karl Marx and Friedrich Engels, Immanuel Kant's complete works, and also the Corpus Coranicum, a documentation of the text of the Koran in its oral and manuscript transmission.

For almost two decades, the BBAW has been committed to the use of digital methods in research in the Humanities and the Social Sciences. In 2001, a digital transformation initiative called TELOTA (The Electronic Life of the Academy)[32] was founded in order to compile, annotate and publish research data and publications in electronic media. Since then, more than 50 comprehensive collections of digital scholarly resources, publications, and critical editions, such as *Travelling Humboldt – Science on the Move*[33] have been made available online under open licences to the research community and the broader public. In addition, TELOTA develops and provides research software for digital editions, such as the editing environment *ediarum* and the web service *correspSearch*. All research data as well as the related software are produced according to well-established standards and best practices from the field of Digital Humanities.

Since its establishment, the Academy has placed particular emphasis on language documentation, including the *Deutsches Wörterbuch* by Jacob Grimm and Wilhelm Grimm, the *Wörterbuch der deutschen Gegenwartssprache* (WDG), and, more recently, the DWDS. In the context of the DWDS, the academy has been engaged in the compilation of large reference corpora including the German Text Archive (hereafter DTA).[34] Since January 1, 2019, the BBAW has been coordinating the Federal Ministry of Education and Research (hereafter BMBF)-funded Centre for Digital Lexicography of the German

[30] http://aaew.bbaw.de/wbhome/Broschuere/index.html
[31] https://www.dwds.de/, see also Klein/Geyken (2010).
[32] The Electronic Life Of The Academy, see https://www.bbaw.de/en/bbaw-digital/telota
[33] https://www.bbaw.de/en/research/alexander-von-humboldt-auf-reisen-wissenschaft-aus-der-bewegung-travelling-humboldt-science-on-the-move
[34] http://www.deutschestextarchiv.de/, see also Geyken et al. (2011).

Language (hereafter ZDL), whose aim is to bring together the scholarly expertise and resources of all historical and contemporary reference dictionaries of the German language from four different academies (BBAW; Göttingen Academy of Sciences and Humanities, hereafter AdWGÖ; Saxon Academy of Sciences and Humanities, hereafter SAW; and Academy of Sciences and Literature Mainz, hereafter AdWMZ) as well as the IDS.

Role: The BBAW will contribute to the Text+ Clusters in all three data domains: it will coordinate the Cluster on *Historical Texts* for the data domain Collections. For the data domain Lexical Resources, the BBAW will coordinate the Cluster *German Dictionaries in a European Context* in collaboration with the IDS and the Competence Center for Lexical Resources at Trier University. It will also contribute to the Cluster *Non-Latin Scripts* in collaboration with the Data Center for the Humanities at the University of Cologne (DCH), a participant in Text+. For the data domain Editions, the BBAW will contribute to the Clusters *Ancient and Medieval Texts* and *Early Modern, Modern, and Contemporary Texts*.

The German National Library (DNB) preserves and provides access to a major part of Germany's cultural heritage in the form of written, pictorial, and sound recordings published in Germany and in German since 1913. The DNB facilitates research projects in a wide range of disciplines by providing the ever-growing digital collection of texts as flexibly as possible. In addition, the DNB offers central services such as the Integrated Authority File (GND)[35], a collaborative platform for standardised, cross-domain machine-readable interlinking of collections and databases.

The DNB is active in the area of library standardisation and coordinates the development of fundamental concepts and standards. It is also one of the main bodies in the German Digital Library (DDB) and is connected to Europeana and within the international Federation of Library Associations (IFLA).

Role: Within Text+, the DNB will manage the data domain *Collections*.

First, the DNB will provide access to the huge and ever-growing collection of 21st century texts. This corpus ranges from contemporary German-language literature and a collection of all daily newspapers to the collection of dissertations from German universities, but also includes scientific articles from German publishers as well as kiosk and consumer literature. Due to the restrictions on access to most of the objects in the DNB's holdings, more flexible access options must be developed based on the legal framework. Together with the scientific community, the DNB will actively participate in the development of a set of derived text formats with reduced information to allow access to protected documents. The DNB will contribute its existing expertise on legal issues and data protection.

Interoperability within the heterogenous collection of the DNB and between other resources will be achieved through the Linked Open Data (LOD) format. One of the tasks within Text+ is the interlinking

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

with other text collections such as the DTA and the targeted inclusion of full texts, from the application of OCR to digitised material from VD16, VD17 and VD18. Here, the competencies in the area of metadata standards form an essential basis. The DNB will actively participate in the development of techniques to link these data with other data sets from Text+. This linking will be realised via authority files such as the GND and via lexical resources such as GermaNet.

The DNB will also bring to bear its leading role in the construction of GND and its expertise and experience in OCR and other digitisation initiatives as well as the knowledge and experiences with LOD in the Cluster *Unstructured Text* for the Text+ data domain Collections. This Cluster will be jointly operated with the SUB.

The DNB will contribute to all Measures in the data domain Collections, but also contribute in Infrastructure/Operations.

The Göttingen State and University Library (SUB) is one of the largest libraries in Germany and a leader in the development of digital libraries. It hosts several digital collections of substantial importance as resources for research in Text+, which are provided by the Göttingen Digitisation Centre. Together with the German National Library (DNB), the SUB manages the specialist department *Library Data*[36] of the German Digital Library and coordinates the activities of DINI-AG KIM[37]. It is the coordinator of DARIAH-DE, member of the National coordinator committee of DARIAH-ERIC and is coordinating CLARIAH-DE together with UniTÜ. The SUB provides a DOI service for the Humanities in co-operation with DataCite, which has already registered over 40,000 data sets, as well as local, national, and international support for the creation of digital editions by an in-house unit[38]. On the international level, the SUB is scientific coordinator of OpenAIRE[39], partner in the European plug-in to the Research Data Alliance[40] (hereafter RDA) (RDA Europe 4.0) and partner in the EOSC-project SSHOC[41] (Social Sciences and Humanities Open Cloud).

Role: The SUB will provide services, data, and competences for all Task Areas. Together with the IDS, the SUB is responsible for the overall coordination of Text+. It leads the Task Area Infrastructure/ Operations and will focus on community services and cross-cutting topics. In particular, the SUB will contribute to the metadata infrastructure in order to increase interoperability and re-usability of the data in Text+. The SUB is part of numerous standardisation committees, such as the TEI-Consortium[42], the Dublin Core Governing Board[43], MODS Editorial Committee[44], the IIIF Consortium, CIDOC CRM-SIG

[36] https://pro.deutsche-digitale-bibliothek.de/fachstelle-bibliothek
[37] *Kompetenzzentrum Interoperable Metadaten*, see https://dini.de/ag/kim/
[38] *Service Digitale Editionen*, see https://www.sub.uni-goettingen.de/digitale-bibliothek/service-digitale-editionen/
[39] https://www.openaire.eu
[40] https://www.rd-alliance.org
[41] https://www.sshopencloud.eu
[42] https://tei-c.org
[43] https://www.dublincore.org/groups/governing-board/
[44] https://www.loc.gov/standards/mods/editorial-committee.html

and LIDO WG. The SUB is significantly involved in the development and advancement of various metadata standards, for instance, by its involvement in the specification of the METS/MODS Application Profile for Digitised Prints, which is the de-facto description standard for material digitised in German libraries.

The SUB will coordinate the Cluster *Early Modern, Modern, and Contemporary Texts* for the data domain Editions and, together with the DNB, the Cluster *Unstructured Text* for the data domain Collections. It hosts the TextGrid Repository and the DARIAH Repository. The latter is part of the thematic service of EOSC-hub and open for ingest and access for all kinds of research data of the Arts and Humanities. The TextGrid Repository is a recognised and valuable resource, in particular for Literary Studies (Philology, Comparative Literature), and will be expanded continuously. It is a CoreTrustSeal certified, community-curated repository and open for the ingest of new data (in different languages). In co-operation with the COST Action "Distant Reading"[45], the ingest of several collections of novels of the 19th and 20th century in six European languages and formatted in TEI-XML is in preparation.

In Collections, the SUB will contribute to the FAIRification of unstructured text by the development of a set of derived formats (according to the research questions formulated by the community and in close collaboration with the DNB and UniTR).

In Editions, the SUB provides generic tools for the creation and publication of digital editions (TextGrid[46], SADE[47], TextAPI[48]) and has extensive experience in teaching digital editing skills and tool usage through training, workshops, and summer schools.

The **North Rhine-Westphalian Academy of Sciences, Humanities and the Arts (NRWAW)**, founded in 1970, is an association of the state's leading researchers and brings together all forms of creative discovery, be they scientific, scholarly or artistic. The NRWAW currently supervises 13 long-term research projects, many of them dealing with textual heritage and with editions in all their aspects, ranging from ancient to modern material and from German language to non-Latin scripts. To ensure state-of-the-art digital methods and to cover the complete project life cycle, the NRWAW has established a central coordination office for Digital Humanities located at the Cologne Center for eHumanities (CCeH) and collaborates closely with the Data Center for the Humanities, Cologne (DCH).

Role: The NRWAW will contribute its long-standing experience in digitisation and in the creation and maintenance of a broad range of different types of editions to the two Clusters *Ancient and Medieval Texts* and *Early Modern, Modern, and Contemporary Texts*. In close collaboration with its coordinating office, the NRWAW contributes ample experience in consultation, in planning and carrying out digital

[45] https://www.distant-reading.net/eltec/
[46] https://textgrid.de
[47] https://gitlab.gwdg.de/SADE
[48] https://subugoe.pages.gwdg.de/emo/text-api/

research projects as well as in data management, archiving and the implementation of sustainability measures. As co-applicant institution, the NRWAW is responsible for the Task Area Editions and will contribute to all Measures within this Task Area (M1–M5).

Text+ is an open and community-driven consortium that **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 of Text+ make important contributions to the Text+ Clusters by providing their research data, services, and expertise. The network of eight Clusters is configured in a modular and scalable fashion. As the data and service portfolio of Text+ will be extended over time, additional Clusters and participating institutions will be added that complement existing strengths of the Clusters. Annual calls for research data and for user stories will be issued that solicit proposals for additional data and for novel ways of utilising and extending the Text+ data and service portfolio.

Individual Participants of Text+ can take on a variety of tasks and roles in Text+: the initial group of individual participants will take up key positions in the governance structure of Text+. Sandra Richter, Ingrid Schröder, and Verena Klemm will act as chairpersons of the Scientific Coordination Committees for the data domains Collections, Lexical Resources, and Editions, respectively, and Vivien Petras will act as initial chairperson of the Operations Coordination Committee (for further details on these roles, see section 3.4.2 below).

Apart from participation in the community-driven governance structure of Text+, individual participants can also contribute their expertise to one or more Clusters as community experts. Individual experts will also be invited to engage in working groups and task forces on specific topics, exchange information, enforce the community network and connect data, methods and people.

Text+ is committed to not only involve established, but also young researchers in both the consortium's activities and governance. By attendance in the plenary, young researchers have the opportunity to participate in the governance of Text+ and to be potentially elected to a coordination committee. Moreover, the data domain-specific Clusters offer PhD students and postdoctoral researchers mentoring and targeted partnerships for projects that centre around the creation and re-use of research data for language and texts. Calls for such partnerships will be issued on a regular basis and an annual community event will provide ample opportunity to present the results of these projects. The consortium also envisages joint teaching events, e.g. at summer schools, as well as joint dissemination events with and for young researchers, e.g. at the large annual conferences of the professional associations.

Professional associations will provide the organisational bridge to the communities of interest in several ways.[49] Many associations have formed working groups for digital methods and data. Members of these working groups will continue and intensify their long-standing dialogue with member institutions of Text+. In addition, professional associations will nominate members for the Scientific and Operations Coordination Committees of Text+ and will be the driving force in the portfolio development of Text+ research data and services.

Furthermore, Text+ coordinates its activities with other stakeholders bridging infrastructure and community, namely with the disciplinary specialised information services (*Fachinformationsdienste*; FID). They cover a considerable part of Text+'s communities of interest and a particular wide range of languages and cultures (see Letters of Support on the Text+ website[50]). The FIDs will not only foster the dissemination of and engagement with Text+'s portfolio in their respective communities, but will also help to involve members of the latter in the further development of the consortium.[51]

[49] see https://www.text-plus.org/en/about-us/academic-societies/
[50] see https://www.text-plus.org/en/about-us/further-partners/
[51] https://www.text-plus.org/en/about-us/further-partners/

3.2 The Consortium within the NFDI

Text+ is institutionally interlinked with numerous NFDI consortia through participation of partners in the following consortia: DataPLANT, FAIRmat, GHGA, KonsortSWD, NFDI4Agri, NFDI4Biodiversity, NFDI4Chem, NFDI4Culture, NFDI4Earth, NFDI4Ing, NFDI4Memory, NFDI4MobilTech, NFDI4Objects, NFDIMatWerk, NFDIxCS, PUNCH4NFDI.

A particularly intensive collaboration was established during several coordination meetings with three other NFDI consortia from the Humanities and Cultural Sciences: NFDI4Culture, NFDI4Memory, and NFDI4Objects. The four initiatives share the conviction that serving their respective communities of interest as well as the NFDI as a whole can best be achieved by means of a jointly coordinated and continuous development which is carried out in complementary fields of action, and by integrating common topics identified in the process into the overarching context of the NFDI.

The interfaces to and the planned collaboration with other applicant Humanities consortia in The collaboration is described in a Memorandum of Understanding (MoU), which is open to be signed by further initiatives. This MoU has been published in a second, updated version[52] in September 2020.[53]

Four cross-cutting topics were identified as joint fields of action in which the aforementioned consortia can not only provide substantial, long-standing expertise but are also prepared to develop common services to the benefit of the NFDI as a whole: (a) metadata, authority data, terminologies, (b) provenance, (c) rights and ethics, and (d) data literacy. In addition, common liaisons are maintained in the context of the *Forum geisteswissenschaftliche NFDI*[54] with the professional associations.

Text+ has also signed both the Berlin Declaration[55] and the Leipzig-Berlin Declaration[56] on NFDI cross-cutting topics and is prepared to enter into close, NFDI-wide knowledge exchange and coordination regarding the Research Data Commons. Text+ will foster the discussion about the Data Commons in the Humanities in order to contribute to the mediation of priorities with other disciplines.

[52] see https://doi.org/10.5281/zenodo.4045000
[53] see Brünger-Weilandt et al. (2020).
[54] see https://nfdi.hypotheses.org/
[55] see Glöckner et al. (2019).
[56] see Bierwirth et al. (2020).

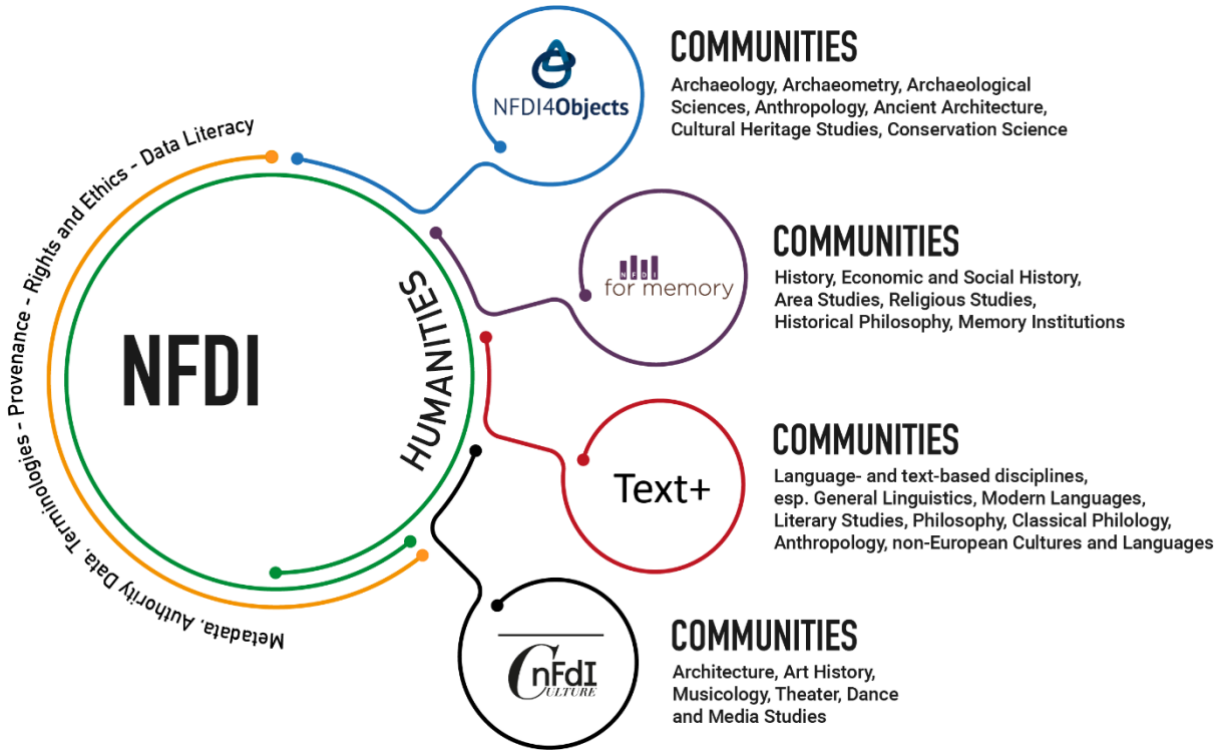

Figure 3.2 NFDI4Culture, NFDI4Memory, NFDI4Objects, Text+ and their communities

Scholarly communication in virtually all disciplines is based on language and text. Therefore, interfaces to cross-disciplinary activities in the field of metadata, authority data, and terminologies with a particular focus on the GND as well as services such as text and data mining are planned. A cooperation on quantitative and qualitative language- and text-based research methodologies will be established with KonsortSWD (*Konsortium für die Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften*), where Text+ focuses on metadata and text resources while KonsortSWD focuses on anonymisation procedures. NFDI4Ing and Text+ will collaborate in scoping out and leveraging the potential of text mining techniques for the extraction of research data from pertinent digital scholarly literature and documents. In the field of language technologies and natural language processing, Text+ will collaborate with NFDI4DataScience on semantic web technologies, text and data mining and methods for transforming unstructured to structured text. With BERD@NFDI, Text+ will cooperate inter alia on text and data mining for speech data and the analysis of (unstructured) text. Text+ has also identified common interfaces with other NFDI consortia as diverse as NFDI4Earth or NFDI4BioDiv.

Generic cross-cutting topics relevant for Text+ and other NFDI consortia beyond the above-mentioned joint topics of the Humanities and Cultural Studies consortia are: data, service and software quality frameworks, search and interoperability solutions, standardisation and certification, as well as technical infrastructure services, particularly for ensuring compliance with FAIR and CARE principles. Technical infrastructure services include authentication and authorisation services, persistent

identifiers for research data, certification services for data centres, as well as easy access to data storage and archiving facilities and high-performance computing.

Text+ is committed to the propagation of common, fundamental RDM strategies, the FAIR and CARE principles, the FAIRification and CAREification of data and, as a prerequisite for this, to the active promotion of the cultural change in all disciplines. Text+ will contribute to cross-cutting topics by sharing expertise and practical experience with other NFDI consortia in the following areas:

Metadata, authority data, terminologies: Text+ will actively promote common reference models and help develop interoperable solutions for metadata application profiles. It will support the usage and enrichment of cross-disciplinary terminologies and authority data, especially regarding the GND.

Provenance: Text+ will promote and help develop solutions for integrating provenance information into data models and curation practices, with inclusion of its manifold aspects which range from technical to legal issues.

Rights and ethics: Text+ will provide expertise and consulting on licensing and contributions to data governance solutions, and it will raise awareness for implicit data bias and associated ethical issues.

Data literacy: At the national and international level, Text+ will organise workshops, training, and summer schools for young researchers on Text+ topics that are of general interest to data-driven research (e.g. text and data mining). It will also provide Open Educational Resources (OER) and contribute to the development of curricula.

Data, service and software quality management and assurance: Text+ will actively contribute to the development of curation criteria and quality standards for research data in the Humanities and related quality management processes, quality management of software and services, and lifecycle management. It will in particular support the CTS[57] certification of data centres.

Search and interoperability solutions: Text+ will offer a federated metadata and data infrastructure. Relevant services include hosting, migration, integration and synchronisation as well as conversion from and to RDF and linking with resources in the Linked Open Data cloud.

Technical infrastructure services: Text+ will offer authentication and authorisation services, persistent identifiers for research data, digital long-term preservation of research data, and development and deployment of Text Data Mining techniques for language- and text-based research data.

3.3 International Networking

Text+ will draw on already established international networks of research data users, research data providers, and research infrastructure initiatives. The applicant and co-applicant institutions contribute to the only two research data infrastructures for the Humanities that have attained landmark status on the current roadmap of the European Strategy Forum for Research Infrastructures

[57] https://www.coretrustseal.org/

(ESFRI) and that have both been established as ERICs (European Research Infrastructure Consortium). Andreas Witt, Head of the Department for Research Infrastructures at the applicant institution, is a member of the Board of Directors of CLARIN ERIC, where he focuses on strategic planning and sustainability. Erhard Hinrichs, the spokesperson of Text+, serves as the National Coordinator for Germany in the National Coordinator Forum of CLARIN ERIC. Nanette Rißler-Pipka, Head of the Digital Humanities group at the co-applicant institution SUB, serves as the National Coordinator for Germany in the National Coordinators Committee of DARIAH ERIC.

Through in-kind contributions for research data management infrastructure services and knowledge sharing, Text+ will be able to create synergies between the national and international levels that will directly benefit Text+ and the two ERICs. More specifically, Text+ will leverage the extensive research data contributed by more than twenty European member countries in CLARIN ERIC and DARIAH ERIC. This will allow Text+ to provide direct links to research data across Europe for the three data domains of Text+. By offering direct access for Humanities scholars in Germany to the national data centres for lexical resources and for reference corpora of other European countries, research barriers can be overcome, for example, and federated search across languages will become possible for the first time. The two ERICs and designated partner institutions of Text+ in Germany are actively contributing to the design and implementation of the European Open Science Cloud (EOSC) through Horizon 2020 projects, e.g. EOSC-Pilot, EOSC-Hub and OpenAIRE-Advance. In OpenAIRE, Wolfram Horstmann, director of the SUB Göttingen, is scientific coordinator and a member of the Executive Board, the steering committee of the pan-European network. In EOSC, services and tools for the discovery, access, use and re-use of resources for advanced data-driven research are built predominantly across disciplines. Additionally, the engagement of Text+ partners in SSHOC will enable a discipline-specific alignment of EOSC, Text+ and the NFDI. In order to facilitate this alignment, Wolfram Horstmann and Andreas Witt will act as liaison partners on behalf of Text+ vis-à-vis EOSC and ESFRI, respectively.

The Union of the German Academies of Sciences and Humanities is a member of ALLEA,[58] the European Federation of Academies of Sciences and Humanities, which currently brings together almost 60 Academies in more than 40 countries from the Council of Europe region. Text+ has direct ties to OPERAS[59], a research infrastructure for the Humanities and Social Sciences via the MWS, which is member of the Executive Assembly of OPERAS and which contributes to Text+ as a participant through its central office in Bonn and its institutes and branch offices in Beijing, London, New Delhi, Rome, Tokyo, Warsaw, Washington. Through the involved libraries, Text+ is also connected to the Digital Humanities initiatives of the national libraries, the Association of European Research Libraries

[58] https://allea.org

[59] Open scholarly communication in the European research area for social sciences and Humanities, see https://operas.hypotheses.org/

(LIBER)[60], the Consortium of European Research Libraries (CERL)[61], and the International Federation of Library Associations and Institutions (IFLA)[62] and incorporates their valuable experiences.

International collaborations of Text+ extend well beyond European partnerships. Text+ partners co-operate on facilitating access to research data with several data centres in the United States, including the CLARIN Centre for Child Language Acquisition Corpora at Carnegie Mellon University, the HathiTrust Research Centre (HTRC), the Linguistic Data Consortium at the University of Pennsylvania, the iSchool at Illinois, and the Language Application Grid at Brandeis University. Text+ partners also maintain strong ties to universities in South Africa via the Global WordNet Association and the South African Centre for Digital Language Resources (SADiLaR). This co-operation focuses on corpus and lexical resources for the indigenous languages of South Africa. Co-operation with South-East Asia and Australia centres around language documentation for endangered languages and involves the Australian Research Council (ARC) Centre of Excellence for the Dynamics of Language[63] and the Center for Endangered Languages Documentation (CELD) in Manokwari (West Papua)[64]. Text+ partners also contribute regularly to the conference and workshop series on Endangered and Lesser Known Languages of India (ELKL).[65]

Finally, member institutions of Text+ are active in the relevant committees of international organisations that advance the standardisation of research data and promote the use of such standards world-wide. These include the RDA[66], the TEI-Consortium[67], the IIIF-Consortium[68], DataCite[69], COAR[70], DC[71], and ISO[72].

3.4 Organisational Structure and Viability

In direct response to **Grand Challenge III**, the governance model of Text+ fosters joint responsibility and consensus building among infrastructure providers and infrastructure users as well as collaboration across disciplinary boundaries. It is summarised in Figure 3.3.

[60] https://libereurope.eu/strategy/digital-skills-services/digitalhumanities/
[61] https://www.cerl.org/
[62] https://www.ifla.org/
[63] http://www.dynamicsoflanguage.edu.au/home/
[64] https://www.celd-papua.org/
[65] https://sites.google.com/site/elkl4agra/tutorials-1
[66] https://www.rd-alliance.org/
[67] https://tei-c.org/
[68] International Image Interoperability Framework https://iiif.io/community/consortium/
[69] https://datacite.org/
[70] https://www.coar-repositories.org/
[71] https://dublincore.org/
[72] https://www.iso.org/; particularly the technical committee ISO/TC37 on Language and Terminology.

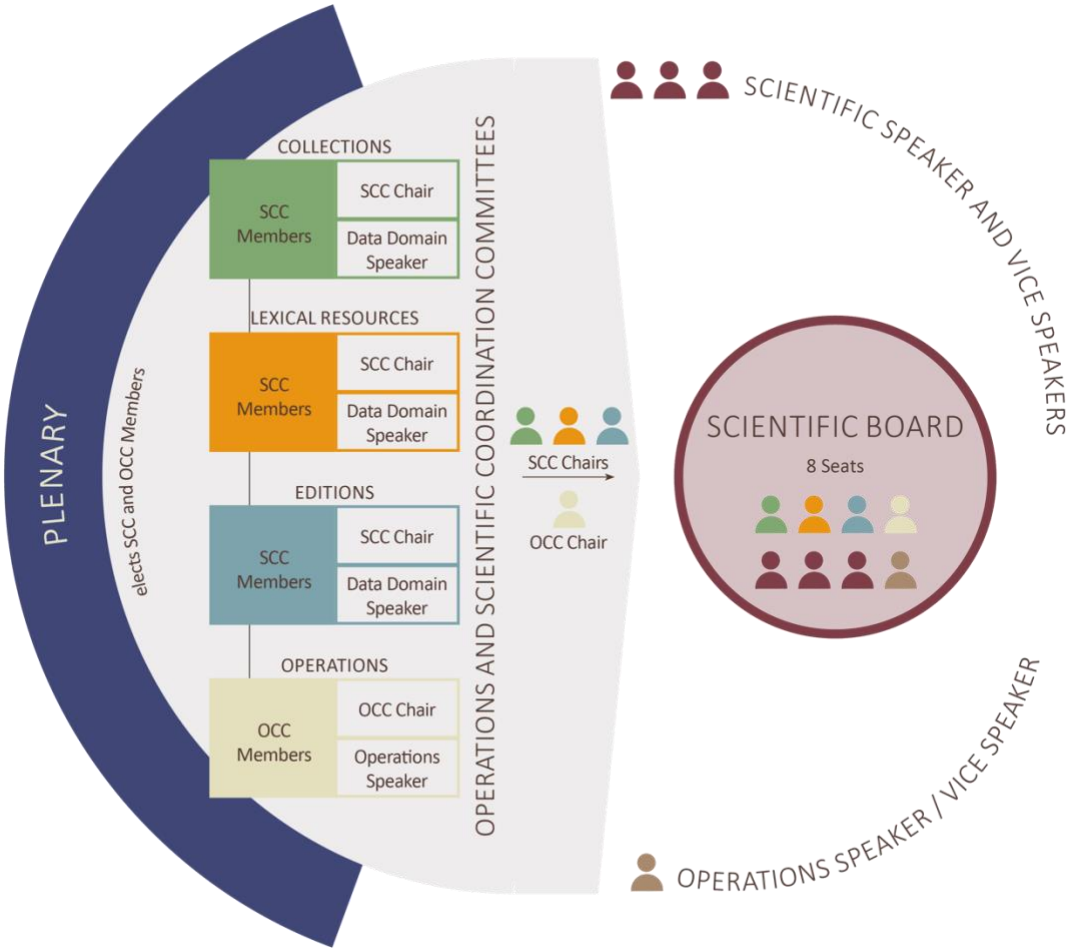

Figure 3.3 Text+ Scientific Governance

3.4.1 Plenary

An annual face-to-face meeting will be the central Text+ community event and will allow for orchestrated decision-making. Representatives of the professional associations (*Fachgesellschaften*), Text+ participants and co-applicants form the plenary, which is announced publicly and open to interested parties. The plenary elects the three Scientific Coordination Committees (SCC), one for each data domain, and an Operations Coordination Committee (OCC), which are at the core of the Text+ governance model. Initially, 20 professional organisations have confirmed to appoint representatives.[73] Moreover, a number of individual experts have expressed their interest to contribute to the SCCs or the OCC.[74] Self-nominations are possible.

[73] https://www.text-plus.org/en/about-us/academic-societies/
[74] https://www.text-plus.org/en/about-us/further-partners/

3.4.2 Coordination Committees

The SCCs and the OCC will continuously evaluate the Text+ portfolio of data, tools, and services and will be instrumental in extending the Text+ portfolio in line with priorities set by the communities of interest. Each SCC is composed of one scientific chair, several scientific members and its data domain speaker, the co-speaker of the correspondent Task Area. The scientific members form a balanced representation of the scientific field with a special focus on professional associations as well as outstanding experts in the given field. The scientific chair of a given SCC is nominated from within the SCC. The data domain speaker cannot become chair of an SCC to prevent conflicts of interest and to guarantee scientific leadership. The OCC is composed accordingly.

The pre-nominated chairs of the SCCs and the OCC are well-recognised and independent researchers (not part of the Text+ funding) and are willing to commit themselves on an honorary basis to serve as acting chairs for the first year. For the Text+ data domains and operations the following persons have been nominated:

- Collections: **Sandra Richter**, director of the German Literature Archive in Marbach, professor and head of the department for Modern German Literature at the University of Stuttgart.
- Editions: **Verena Klemm**, professor of Arabic and Islamic Studies at the University of Leipzig and head of the long-term project *Bibliotheca Arabica. Towards a New History of Arabic Literature*.
- Lexical Resources: **Ingrid Schröder**, professor of Low German Language and Literature, head of the department at the University of Hamburg and head of the project *Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)*.
- Operations: **Vivien Petras**, professor for Information Retrieval and head of the Berlin School of Library and Information Science at Humboldt Universität zu Berlin.

The members of the four coordination committees for the first year will be nominated at the beginning of the funding period in co-operation with the professional associations, who will propose candidates out of their community. During the first Text+ Plenary, elections for all coordination committees and their chairs will be held. The SCCs/OCC review the services and give concrete recommendations to the Scientific Board regarding proposed projects, new participants and the development of the data portfolio.

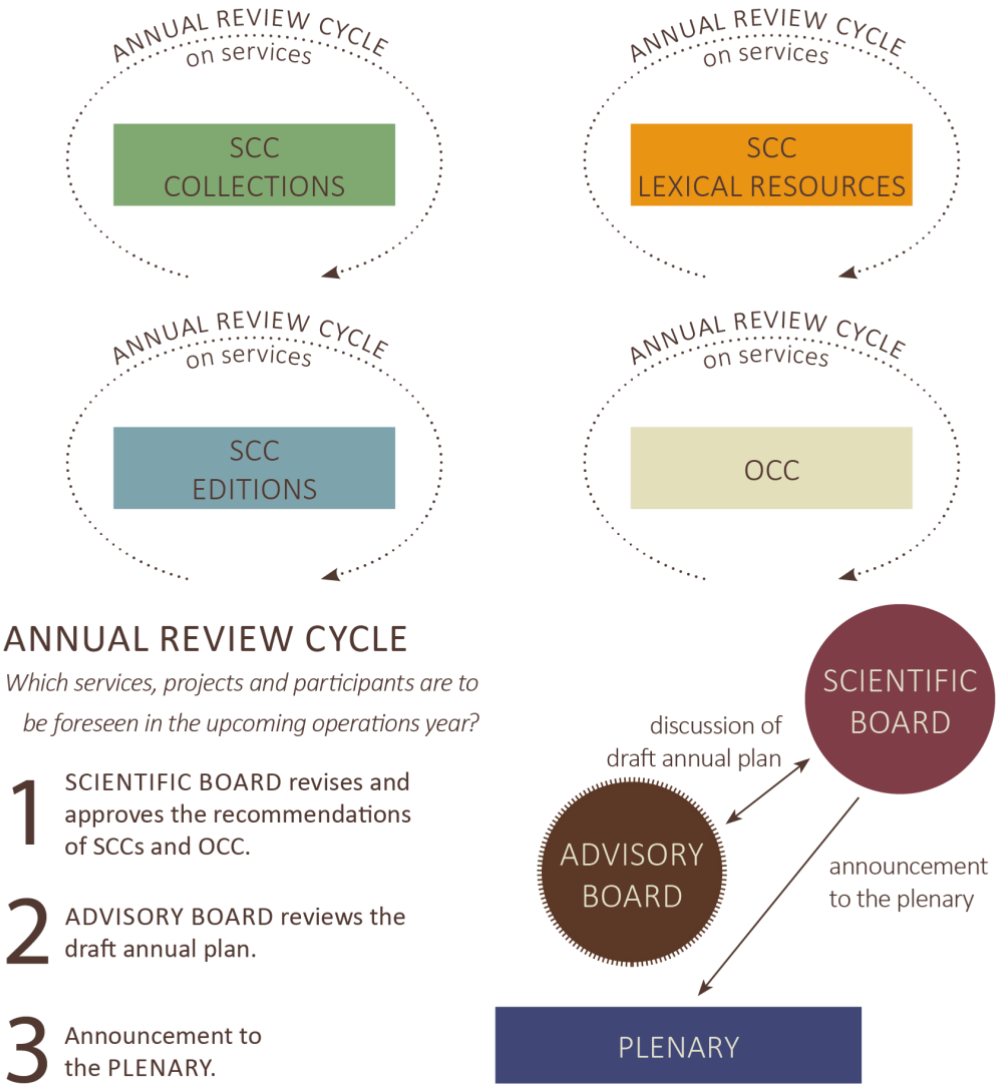

Figure 3.4 Annual review cycles of SCCs, OCC and the Scientific Board

3.4.3 Scientific Board and Institutional Embedding

The **Scientific Board** is composed of the four chairs of the SCCs and the OCC, the scientific speaker and the two vice-speakers, and the operations speaker. No member of the Scientific Board has a deputy, with one exception: in case of unavailability of the operations speaker, their vice-speaker is authorised to act as appointed representative.

For the scientific speaker, Erhard Hinrichs, and the operations speaker, Regine Stein, the following vice-speakers are appointed:

- Scientific vice-speaker: **Andrea Rapp**, professor of German Studies, Computer Philology and Medieval Studies at TU Darmstadt, former vice-president of TU Darmstadt, and second chairperson of the TextGrid Association.

- Scientific vice-speaker: **Elke Teich**, professor of English Linguistics and Translation Studies at Saarland University, spokesperson of the Collaborative Research Centre SFB 1102 and head of the Saarbrücken CLARIN-D centre.
- Operations vice-speaker: **Andreas Henrich**, professor of Media Informatics, Faculty of Information Systems and Applied Computer Sciences at the University of Bamberg.

The Scientific Board revises and approves the recommendations of the SCCs/OCC and elaborates the draft annual plan for discussion with the Advisory Board. The Advisory Board consists of representatives and scientifically recognised members and meets in conjunction with the Text+ Plenary. Final decision-making on the annual plan lies with the Scientific Board.

The Scientific Board as well as SCCs and OCC meet at least quarterly, following an annual cycle: the first meeting is for the annual agenda setting, the second meeting for gathering possible changes of the portfolio, the third meeting for issuing recommendations and the fourth meeting for review and debriefing.

Scientific speakers, data domain speakers and operations speakers form the **Administrative Board**, which is responsible for the implementation of the work programme including technical and financial monitoring of the ongoing work. All members hold permanent positions at their institutions and have been assigned by their home institutions to take on leading roles in Text+ for the entire period of the current construction phase of the NFDI. The Administrative Board meets on a regular monthly basis.

The **Institutional Board** consists of the heads of the (co-)applicants (Christiane Dusch, NRWAW; Martin Grötschel resp. Christoph Markschies, BBAW; Wolfram Horstmann, SUB; Henning Lobin, IDS; Frank Scholze, DNB) and contributes in overarching and strategic questions, in particular with respect to sustainability and co-operation with the NFDI association. It meets once per year and can be summoned upon request.

The scientific speaker and the operations speaker coordinate the consortium and all its activities and are supported by the Text+ office, which will help to ensure maximum transparency of decision-making in Text+ by preparing minutes and publishing all relevant documents. The work of all Text+ committees will be governed by a set of by-laws. Consensual decision-making is aimed for in all committees and boards, but decisions at a quorum of 2/3 with one vote by each member are possible. Votes are personal and non-transferable. In case of conflict, final decisions are taken by the Institutional Board.

3.4.4 Viability

The viability of the organisational structures throughout the initial five-year funding phase of the NFDI will be ensured by the strong commitment of the (co-)applicants and all participants of Text+. The Letters of Commitment (see Appendix **Error! Reference source not found.**) include substantial allocation of institutional funds and provisioning of research data and services.

All Text+ (co-)applicants are decidedly institutions at the interface of infrastructure and academia with a respective mandate and long-standing experience in developing and maintaining research infrastructure services. Due to their long-term institutional funding, they are all in a strong position to align future developments of their own organisational structures and staffing plans with the needs of their communities of interest within the NFDI. They commit themselves to systematically evaluating and integrating these needs into their strategic planning and to allocating staffing levels in accordance with such continuous evaluations. In order to ensure the best possible harmonisation of such processes with the overall developments taking place within the NFDI, a dedicated task in Measure *Co-operations within the NFDI: Sustainability* (see section 5.5.1) is included into the Text+ work programme.

If NFDI funding is discontinued, Text+ (co-)applicants will jointly have in place an exit strategy in the form of an operating agreement. The agreement will include data preservation and accessibility services and will address continued availability of Text+ data and services in accordance with the findings of continuous community-driven evaluations. This sustainability model of Text+ will be fully integrated with the overarching sustainability efforts of the NFDI, which are to be coordinated by the NFDI Directorate. The operating agreement of Text+ will be open and scalable: all present and all future participants of the Text+ consortium as well as other interested parties are invited to join.

In addition, Text+ applicant institutions will promote a general NFDI business model based on the open access model. The Text+ consortium is convinced that scientific data that are the result of publicly funded research must be made available to the public and especially to researchers free of charge if new opportunities for research are to be created in the long term and if an essential contribution to the preservation of the associated cultural heritage is to be made possible. To this extent, it must be possible to structurally embed funding organisations and researchers' institutions in financing the fees of the NFDI. Here, in close consultation with the NFDI Directorate, efforts should be made to establish appropriate funding instruments and programmes at the national and European level in the same way as it is currently becoming common practice to raise a flatrate Open Access publication fee to cover the costs of publishing research results in Open Access journals.

3.5 Operating Model

The operating model of Text+ relies on the Text+ office and the close interaction between its two components: its Scientific Office, led by the Scientific Speaker, and its Infrastructure/Operations Office, led by the Infrastructure/Operations Speaker. The Text+ office is responsible for all matters of documentation and reporting. It will provide information and outreach communication for the communities of interest and will support all activities that pertain to the Text+ plenary.

The Scientific Office will co-operate closely with the three SCCs and will offer administrative support to the SCC Chairs. The Infrastructure/Operations Office will co-operate closely with the OCC and will offer administrative support to the OCC Chair. Together, Scientific Office and Infrastructure/

Operations Office will ensure that the information flow and organisational support are available to all participating parties. Stable and regular management structures will be in place to ensure the targeted coordination of Text+. The full range of Measures undertaken by the Text+ office is detailed in Section 5.5: Administration, Collaboration and Sustainability.

The Text+ office will be instrumental for the effective communication of user requirements, which will be determined in the SCCs. To ensure that the user communities of Text+ determine the community requirements of the Text+ research data infrastructure, the SCCs are composed of members that represent Text+'s communities of interest.

The operating model of Text+ will comply with the non-profit/public benefit requirements and with the German Value Added Tax Act. Unless clearly identified as subcontracts, all exchanges of services among (co-)applicants and/or participants within Text+ or with other NFDI consortia will not involve financial transactions of any kind and will fulfil the criteria of being for the common good (*gemeinnützig*). Consequently, such exchanges of services will not be subject to German Value Added Tax (*Umsatzsteuer*). The terms of the operating model of Text+ will be codified in a Cooperation Agreement, which will be signed by all (co-)applicant and participant institutions involved in Text+. For the time being, Text+ does not plan to establish a legal entity of its own. Rather, it will closely co-operate with the NFDI Directorate and will look to the Joint Science Conference (GWK)[75] and to the NFDI Directorate for guidance as to how the NFDI as a whole will best be set up as a legal entity. Text+ fully supports the ongoing plans to establish a registered non-profit organisation (*eingetragener, gemeinnütziger Verein*) for the NFDI, but it is open to alternative proposals for effective governance and operating models of the NFDI as a whole. Text+ is able to build on a broad range of infrastructure components, which will be described in the next section in the context of its RDM strategy.

[75] https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/gwk-agreement-engl.pdf

4 Research Data Management Strategy

The RDM strategy defines the fundamental path towards the integration of data, applications, tools, and services into an interoperable, standards-based infrastructure that adheres to the FAIR and CARE principles (see section 4.3). Moreover, the RDM strategy specifies fundamental processes to maintain and evolve this infrastructure in close collaboration with both the communities of interest and the NFDI in general.

Section 4.1 *State of the Art and Needs Analysis* outlines an analysis of the requirements for research data management in Text+ and its envisaged state after the first phase of Text+ funding. This outline is based on the state of the art of research data management in the Humanities and on the consortium's expertise and experience. Section 4.2 *Metadata Standards* describes the approach of Text+ towards the usage and evolution of metadata. Section 4.3 *Implementation of the FAIR Principles and Data Quality Assurance* explains the way Text+ implements the guiding FAIR and CARE principles and ensures data quality across all its offers and services. Section 4.4 *Services Provided by the Consortium* summarises the services provided by the Text+ consortium.

4.1 State of the Art and Needs Analysis

Needs Analysis. The RDM strategy of Text+ needs to master the diversity of research data and the diversity of its communities of interest (see section 2.1, **Grand Challenge I**). These diversity requirements can only be satisfied by an RDM strategy that is able to integrate new communities of interest and additional data providers and data domains. Such openness can best be accommodated by a research infrastructure whose design and technical implementation is scalable and adaptable. Text+ will adopt the architecture of a geographically distributed network of thematic clusters (see section 3.1), which is particularly suitable for the integration of additional participants in the context of future NFDI development, but also motivated by constraints on intellectual property rights.

Additional needs for the RDM strategy were solicited from the communities of interest through the following three instruments:

1. community workshops,
2. a call for user stories and
3. a call for data.

The lessons learned from the community workshops[76] reinforced the expectation by the communities of interest that Text+ should exhibit disciplinary diversity. The fact that more than 120 user stories were submitted to the initial call for participation corroborates the high demand and the timeliness of

[76] http://www.forschungsinfrastrukturen.de/doku.php

a research infrastructure for language- and text-based disciplines. Likewise, the responses to our call for data provide a solid basis for further community-driven portfolio development.

Another factor that makes a concise national RDM strategy for language and text data an absolute necessity is that respective research data are generated in thousands of small-scale projects, as described in **Grand Challenge I** (see also the overarching Objective 1).[77] At the other end of the spectrum of research data providers, there is a wealth of research data that is provided by national hubs such as the DNB and other research libraries, as well as by coordination actions such as the DFG-funded OCR-D initiative (*DFG-funded Initiative for Optical Character Recognition Development*)[78]digitised collections into machine-readable full text format. Once such collections are available in full text, the integration of their materials into the Text+ research data portfolio will become an ongoing task. The wide spectrum of data providers requires a coordinated and comprehensive RDM strategy. Text+ is ready to address and put into practice such a strategy in close coordination with research data providers and with the communities of interest that participate in Text+.[79]

Consortium-Specific Experience and Contributions to the State of Art. Text+ institutions have in-depth experience in key areas of research data management and of digital research infrastructures at their disposal. This experience has been built up over many years and has been consolidated in joint research infrastructure initiatives – most recently in the CLARIAH-DE consortium.

- **Distributed network of certified data repositories**

The member institutions of Text+ have developed several mature data repository solutions that have already been in operation for many years already and that are being used for the curation and depositing of Humanities research data.

The network of CLARIN-D repositories[80] and the TextGrid Repository[81] have been certified with the CoreTrustSeal, an elaborated and standardised process to promote trust and confidence in the shared data resources.[82] The certification of the DARIAH-DE Repository[83] is in progress. All these repositories provide a secure and citation-enabled platform for different types of Humanities data. For TEI text-specific data and editions, the TextGrid Repository and the DTA Base Format[84] with accompanying data ingest[85] and data quality tools[86] offer specialised functions and unique virtual research environments

[77] see https://gepris.dfg.de/gepris/ for DFG-funded projects and similar portals maintained by other national funding agencies and research foundations.
[78] http://www.ocr-d.de
[79] see https://www.text-plus.org/en/about-us/academic-societies/
[80] https://www.clarin-d.net/en/preparation/find-a-clarin-centre
[81] https://textgridrep.org/
[82] CoreTrustSeal, https://www.coretrustseal.org/
[83] https://repository.de.dariah.eu/publikator/
[84] http://www.deutschestextarchiv.de/doku/basisformat/
[85] http://www.deutschestextarchiv.de/dtae/submit/clarin
[86] http://www.deutschestextarchiv.de/doku/dtaq

for editing and collaborating on textual data sources. For OCR and structural detection, the OCR-D initiative offers specialised tools and reference data from the 16th to the 19th century. TELOTA[87] provides a large-scale hosting infrastructure for digital editions and digital academy projects, including user support.

- **Federated metadata and content search (FCS)**

In a distributed infrastructure, a federated search for data and tools that are housed in different repositories provides a significant added value for scholars who otherwise would have to execute separate search queries for each repository. CLARIAH-DE already provides such a federated search. The federated search options range from the CLARIN-D approach of metadata-based search such as the Virtual Language Observatory (VLO) and Federated Content Search (FCS, based on standardised SRU/CQL search protocols) to the DARIAH Generic Search based on the Data Modelling Environment (DME).[88] While different kinds of data and requests require different search strategies, it still is an open issue how search in distributed and heterogeneous data can be conceptually unified and technically merged.

- **Standardisation of technical infrastructure components (AAI, PID, metadata infrastructure)**

Text+ institutions have developed a mature authentication and authorisation infrastructure (AAI), a persistent identifier (PID) service, and metadata infrastructure services that are in operation in CLARIAH-DE. These will be re-used and further developed in Text+. The services are described in more detail in sections 4.2 and 4.3 below.

- **Standardisation of data encoding**

Text+ partners have played a leading role in developing guidelines such as the DFG *Recommendations for Data Standards and Tools for Language Corpora*. The DTA Base Format (DTABf) has provided a major contribution to the state-of-the-art for the encoding of historical corpora and for the TEI community in particular. Further contributions to the standardisation for language and text resources have been made by individual experts through their involvement in relevant standardisation committees.

- **Metadata, authority data, terminologies**

Text+ can build on unique experiences in the development and maintenance of metadata standards, application profiles, and authority data such as the GND, provided by the library partners in Text+.[89]

[87] http://www.bbaw.de/telota/telota
[88] https://clariah.de/en/re-use-data/find-data
[89] https://www.dnb.de/EN/Professionell/Standardisierung/Standardisierungsausschuss/standardisierungsausschuss_node.html, https://dini.de/ag/kim/

Community-specific contributions such as TaDiRAH[90], a multilingual taxonomy of digital research activities in the Humanities, also play an important role.

- **Community involvement, training and consulting**

Training measures are provided and carried out by Text+ partners through activities such as nestor,[91] FOSTER Plus,[92] and OpenAIRE,[93] as well as the strategic partnership of CLARIN and DARIAH with the European Summer School in Digital Humanities (ESU[94]). Consulting services focus on best practices for research data management, construction of data management plans, and hosting services for research data. The large number of user inquiries to the CLARIAH-DE helpdesk underscores the high demand for such consulting services. Community involvement further extends to institutionalised interaction with professional associations in the Humanities. They have reacted to the new opportunities and challenges by forming working groups on digital research data and research methods. Scholars participating in Text+ also take an active role in the working groups on digital research data and research methods. These working groups play a crucial role in the promotion of digital research data and methods in the communities of interest for Text+.

Text+ partners have gained this experience in diverse aspects of research data management and community involvement through their participation in numerous national and international initiatives. Text+ partners have provided research data management solutions for several DFG Collaborative Research Centres and Priority Programmes (see section 2.2). Member institutions of Text+ have established a close co-operation with all BMBF-funded e-Humanities centres that focus on text- and language-based data and participate in BMBF-funded projects for quality management.

Another major initiative that members of the Text+ consortium have participated in for many years is the DOBES[95] initiative. It created a highly regarded and technologically advanced data infrastructure for multi-modal data of endangered languages from different language families and different parts of the world. The CLARIN Knowledge-Centre for Linguistic Diversity and Language Documentation offers RDM support for ethnologists and linguistic typologists who engage in fieldwork for language documentation around the world. Consulting services also extend to ethical and legal issues arising in the context of RDM.[96]

---

[90] https://vocabs.acdh.oeaw.ac.at/tadirah/, GitHub: https://github.com/dhtaxonomy/TaDiRAH
[91] Network of expertise in long-term storage of digital resources https://www.langzeitarchivierung.de/Webs/nestor/DE/Home/home_node.html
[92] "Fostering the practical implementation of Open Science in Horizon 2020 and beyond", see https://www.fosteropenscience.eu/
[93] Open Access Infrastructure for Research in Europe, https://www.openaire.eu/
[94] https://esu.fdhl.info/
[95] http://dobes.mpi.nl/, see also Wittenburg et al. (2002).
[96] https://www.clarin-d.net/en/training-and-helpdesk/legal-helpdesk, see Kamocki et al. (2018).

4.2 Metadata Standards

Creation and harvesting of metadata in Text+ will proceed on the basis of widely accepted metadata standards and domain-specific metadata standards for language- and text-based research data. Metadata harvesting from distributed data repositories follows the widely used OAI-PMH standard.[97] This Text+ metadata infrastructure of metadata standardisation and metadata harvesting has been jointly developed by the infrastructure projects CLARIN-D and DARIAH-DE. It provides a solid foundation for the metadata infrastructure of Text+.

The OAI-PMH protocol requires that metadata is provided according to the standard of the DCMI[98], called oai_dc[99]. Other library-centred data formats in use are MARC 21[100], MODS[101], METS[102], PREMIS[103], and the archival standard EAD[104]. In addition to these metadata formats, Humanities scholars and projects use other formats in their research data management. These formats include customisations of TEI-P5[105] (the header part) based on the guidelines of the TEI as a widely used document file format; ontology-based formats such as schema.org[106], SKOS[107] or EDM[108] and extensions; proprietary application profiles based on CIDOC CRM (ISO 21174) and compatible models.[109] Extensions of Dublin Core for language related research data are available with the Open Language Archive Community's specification (OLAC)[110]. Other repositories follow international standards (ISO) for representing metadata for language resources. The underlying model is standardised in the form of an international standard in ISO 24622-1: *Language Resource Management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model*.[111] This model is instantiated by ISO 24622-2, which is referred to as *Part 2: Component Metadata Specification Language*. Tools that make use of CMDI have been developed for the complete FAIR process, including referencing services such as the VLO[112]. These tools make the data findable and accessible, providing essential information on the data for tools and thus fostering interoperability and re-usability. Each of these metadata standards is already used in data repositories containing many thousands of digital objects (for example, over 70,000 sets of research data are described in CMDI in German institutions

[97] Open Archives Initiative Protocol for Metadata Harvesting, https://www.openarchives.org/pmh/
[98] Dublin Core Metadata Initiative, https://www.dublincore.org/specifications/dublin-core/dces/
[99] http://www.openarchives.org/OAI/2.0/oai_dc.xsd
[100] Machine-Readable Cataloging, see http://www.loc.gov/marc/marcdocz.html
[101] Metadata Object Description System, http://www.loc.gov/standards/mods
[102] Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/
[103] Preservation Metadata: Implementation Strategies, https://www.loc.gov/standards/premis
[104] Encoded Archival Description, http://www.loc.gov/ead/
[105] https://tei-c.org/Guidelines/P5/
[106] https://schema.org/
[107] Simple Knowledge Organization System, https://www.w3.org/2004/02/skos/
[108] Europeana Data Model, https://pro.europeana.eu/page/edm-documentation
[109] http://new.cidoc-crm.org/get-last-official-release
[110] http://www.language-archives.org/OLAC/metadata.html
[111] CMDI, https://www.iso.org/standard/37336.html and https://www.clarin.eu/content/component-metadata
[112] http://vlo.clarin.eu/, see also van Uytvanck et al. (2010).

alone). Given the variety of tools, a mapping strategy has been tested. In DARIAH-DE, a general mapping environment, the DME,[113] has been created to help metadata administrators map the data structure of one metadata format to another, creating roundtrips for conversion.

For uniquely referencing research data, PIDs are used. Most data repositories make use of Digital Object Identifiers (DOI), Handles, or Uniform Resource Names (URN). The handle service is collectively provided by the Persistent Identifiers for eResearch (hereafter ePIC), whose members include partners of Text+. With these PID services, the partners implement the international standard ISO 24619:2011 *Language Resource Management – Persistent Identification and Sustainable Access (PISA)*.

The aforementioned standards constitute the fundamental core set of standards supported by Text+. All of them are either maintained by international standards developing organisations such as ISO or are internationally accepted de facto standards as, e.g., those maintained by the Library of Congress including MARC21, METS, or PREMIS. Text+ partners actively participate in evolving these (de facto) standards by contributing to the respective standards developments. To achieve this, each Task Area related to a research data domain comprises the Measure M3 *Standardisation Activities*, which manages and coordinates the Text+ standardisation activities governed by the respective objectives.

The minimum set of metadata necessary is defined by the Dublin Core elements used in oai_dc, of which a subset will be implemented as mandatory for Text+ search and retrieval. In addition, domain-specific target metadata schemas will be selected according to the specific needs of the addressed data domains and their community standards, as no single (meta-) data format fits all requirements. A respective list of established and recommended target schemas will be maintained and conversion to these will be particularly supported by the Text+ metadata infrastructure. Moreover, conversion from and to RDF and linking with resources in the Linked Open Data cloud will be offered by reusing existing standard models and terminologies. Text+ will contribute to the development of NFDI-wide ontology-based solutions with a specific focus on combining metadata formats with authority data or terminology recommendations. The maintenance of the domain-specific languages, the provision of the required tools and processes to ensure interoperability, and the creation of guidelines and best practices are implemented through Measure M3 Interoperability and Re-usability in the Task Area Infrastructure/Operations.

4.3 Implementation of the FAIR and CARE Principles and Data Quality Assurance

While measures for the implementation of the FAIR principles and Data Quality Assurance are at the core of any NFDI, we also elaborate here on ethical and legal issues involved in data collection referring to the CARE principles.[114]

[113] see Gradl/Henrich (2014).
[114] CARE Principles for Indigenous Data Governance, see https://www.gida-global.org/care

4.3.1 Implementing FAIR and CARE

Text+ will put the FAIR principles into concrete practice with the following approach: the Task Area Infrastructure/Operations offers generic services that constitute the FAIR foundation of the Text+ service portfolio. Each data domain and its Clusters (see sections 3.1 and 4.4) adapt these services to the needs and research priorities of the specific communities of interest, including focused community activities and community involvement. Thereby, the different backgrounds and levels of experience in each community of interest with regard to FAIR data management and FAIRification can be taken into account in the best possible way. Moreover, the FAIR principles are thoroughly incorporated into data selection criteria and standardisation activities. An important building block for establishing a network of trust is a certification procedure for data repositories that is based on accepted international standards. Therefore, Text+ will require a certification of FAIRness from each data repository that wants to become part of the Text+ federated research data infrastructure. Preferably, repositories should become certified according to the CoreTrustSeal (CTS)[115] requirements and processes, when entering the Text+ network, and will be supported in the certification process.

The following table provides an overview of the Text+ Measures in the Task Area Infrastructure/Operations (see section 5.4) and respective data domain Measures for implementing the FAIR guiding principles. By adapting the services of Infrastructure/Operations, the data domains will automatically follow the respective standards. The detailing of the FAIR principles is based on the original version of FAIR:[116]

FAIR Guiding Principles	**Text+ Measures**
F1. (meta)data are assigned a globally unique and persistent identifier	M2 Accessibility in Task Area (TA) 5.4: ePIC and DOI service
F2. data are described with rich metadata (defined by R1 below)	M1 Reference Implementation in TA 5.1-5.3: data curation, data ingest, FAIRification.
F3. metadata clearly and explicitly include the identifier of the data it describes	M1 Reference Implementation in TA 5.1-5.3: data curation, data ingest, FAIRification.
F4. (meta)data are registered or indexed in a searchable resource	M1 Findability in TA 5.4: registries, federated metadata and content search
A1. (meta)data are retrievable by their identifier using a standardised communications protocol	*See A1.1-A1.2*
A1.1 the protocol is open, free, and universally implementable	M2 Accessibility in TA Infrastructure/Operations: ePIC and DOI service using http(s)
A1.2 the protocol allows for an authentication and authorisation procedure, where necessary	M2 Accessibility in TA Infrastructure/Operations: authentication and authorisation infrastructure using the Shibboleth protocol
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	M3 Interoperability and Re-usability in TA Infrastructure/Operations: standardisation and best practices and M3 Standardisation in the TA of the data domains

[115] https://www.coretrustseal.org/
[116] https://www.nature.com/articles/sdata201618

I2. (meta)data use vocabularies that follow FAIR principles	M3 Interoperability and Re-usability in TA Infrastructure/Operations: Metadata, Authority Data, Terminologies service including GND Agency and M1 Reference Implementation in in the TA of the data domains
I3. (meta)data include qualified references to other (meta)data	M3 Interoperability and Re-usability in TA Infrastructure/Operations: Metadata, Authority Data, Terminologies service including GND Agency and M1 Reference Implementation in in the TA of the data domains
R1. meta(data) are richly described with a plurality of accurate and relevant attributes	*See R1.1-R1.3*
R1.1. (meta)data are released with a clear and accessible data usage license	M3 Interoperability and Re-usability in TA Infrastructure/Operations: standardisation and best practices and M1 Reference Implementation in in the TA of the data domains, M2 Portfolio Development in TA Collections in particular
R1.2. (meta)data are associated with detailed provenance	M3 Interoperability and Re-usability in TA Infrastructure/Operations: standardisation and best practices and M1 Reference Implementation in the TA of the data domains
R1.3. (meta)data meet domain-relevant community standards	M3 Interoperability and Re-usability in TA Infrastructure/Operations: standardisation and best practices and M3 Standardisation in the TA of the data domains

It is now widely accepted that the FAIR data-sharing standard is not fully adequate when dealing with fieldwork data collected from endangered and, more generally, minority language communities. In this context, the protection of interests and expectations of the members of these indigenous communities requires pro-active action from Western infrastructure providers.[117] Such efforts are guided by the CARE principles (Collective benefit, Authority to control, Responsibility and Ethics). These principles, which will be embedded in Text+, are purpose- and human-oriented; they are aimed at guaranteeing that value extracted from indigenous data (such as endangered language data) is explored in ways compatible with the interests and cultural requirements of indigenous communities. Data centres in Text+ which handle linguistic and cultural data from indigenous communities will reach out to these communities with two major goals. On the one hand, they will provide for means (including visits to the centres and/or the communities) that will allow communities to control the way their data are represented, not only with regard to access level protocols, but more generally with regard to all aspects visible to users accessing a given collection via the net. This also includes the option to provide for a specific community access (portal) to the data, using languages, technologies and layouts accessible to community members. On the other hand, data centres in Text+ will actively engage in providing technologies and knowledge to indigenous communities and to partner institutions in less developed countries with the aim of allowing them to run their own data infrastructures and thus to retain full, and if they wish, exclusive control over their data.

[117] see Kukutai/Taylor (Eds.) (2016).

Another ethical challenge that Text+ will have to face is the avoidance of information bias. In language science, biased results may usually be attributed to the imbalance of the input data (corpus).[118] However, balance can hardly be assessed and achieved *in abstracto*; rather, a corpus that is perfectly balanced from the point of view of a specific research question will no longer be balanced enough when it is re-used in another project.[119] To mitigate this in a constantly evolving, multidisciplinary environment such as Text+, a form of dynamic (or *bespoke*) selection of data for a specific query is required. In Text+, this can be achieved via extensive and detailed metadata, which allow compiling balanced sets of training data for machine learning via a dynamic selection of data according to a set of precisely-defined criteria. Moreover, in cases where bias in training data cannot be fully avoided *ex ante*, the metadata used in Text+ will allow tracing the results back to the underlying data and to assess potential bias *ex post* by consulting the documentation of the data used for training.

The ethical issue of data privacy has its legal counterpart in personal data protection. Alongside a very broad definition of personal data, which necessarily includes a large portion of Text+ data, the General Data Protection Regulation lists seven principles relating to the protection of such data: 1) lawfulness, fairness and transparency, 2) purpose limitation, 3) data minimisation, 4) accuracy, 5) storage limitation, 6) integrity and confidentiality and 7) accountability.[120] The imperative of Privacy by Design requires that appropriate technical and organisational measures to safeguard these principles be proactively implemented both at the conception phase and at the time of the processing itself. In the delicate context of language data, this requires a thorough analysis of associated risks for individuals (in a Data Protection Impact Analysis), and careful implementation of techniques such as Select before you Collect, or Functional Separation.[121] Text+ will provide guidelines and assistance on these aspects.

4.3.2 Data Quality Assurance

An essential component of any data management strategy concerns the data quality and the quality of associated data services. Quality assessment of research data has to address different levels of compliance. Ultimately, the most crucial aspect of research data quality lies in the correctness and veracity of its contents. At the same time, this aspect of data quality is the hardest to assess and can only be determined by strict peer review. However, as cases of academic misconduct have shown, even peer reviewing cannot provide an absolute safeguard. Most, if not all, of the research data that will provide the initial data portfolio of Text+ are the result of (long-term) research projects that were

[118] see Diesner (2019).
[119] see Kupietz (2015).
[120] Principles relating to processing of personal data, Art. 5 GDPR.
[121] For the general description of Privacy by Design, see: Cavoukian (2009); for a specific application of this requirement to language data, see: Kamocki/Stauch (2020).

subject to stringent peer review. The same holds true for the publications contributed by the DNB and SUB, which have been vetted by major publishing houses or by the scientific community.

As Text+ will expand its portfolio in close interaction with and through the joint planning by the SCCs and the OCC, quality assessment of research data will be an ongoing task in Text+ and the FAIR and CARE principles as well as quality are two of five main data selection criteria for the portfolio development (see section 5). Data and metadata quality assurance of Text+ will build upon a (meta-)data quality analysis framework for datasets based on different Cultural Heritage metadata schemas: this assessment tool,[122] developed at and hosted by Text+ partner GWDG, was created together with the Europeana Data Quality Committee[123] and is used by Europeana for its platform with approx. 50 million records, as well as for several library catalogues. It is adaptable to different metadata schemas, designed to analyse small and big data and has a data analysis backend, and a reporting interface helping data curators to form a data improvement strategy. In addition, Text+ takes up on current work in the BMBF-funded project KONDA[124], which is developing a continuous quality management process for dynamic research data, differentiating according to data, data models and data transformations over the entire lifecycle of data.

Reliability checks of tool and service output, especially for automatic search and annotation tools, need to be based on accepted measures of recall and precision. Since the size of the data sets that search and annotation tools process tends to be very large, data quality checks need to be performed by random sampling methods that can be evaluated by human experts. Such evaluations can then be used to further improve the quality of tools and services.

Quality assessment is by no means limited to research data alone, but also pertains to software tools and services. Quality assessment of the latter, therefore, constitutes an important aspect of the RDM strategy of Text+ and is addressed in Measure M2 Accessibility of Task Area Infrastructure/Operations e.g. through support of CTS certification of repositories, monitoring services, and additional web analytics.

[122] see Király (2017), http://pkiraly.github.io/
[123] https://pro.europeana.eu/project/data-quality-committee
[124] https://www.sub.uni-goettingen.de/en/projects-research/project-details/projekt/konda/

4.4 Services Provided by the Consortium

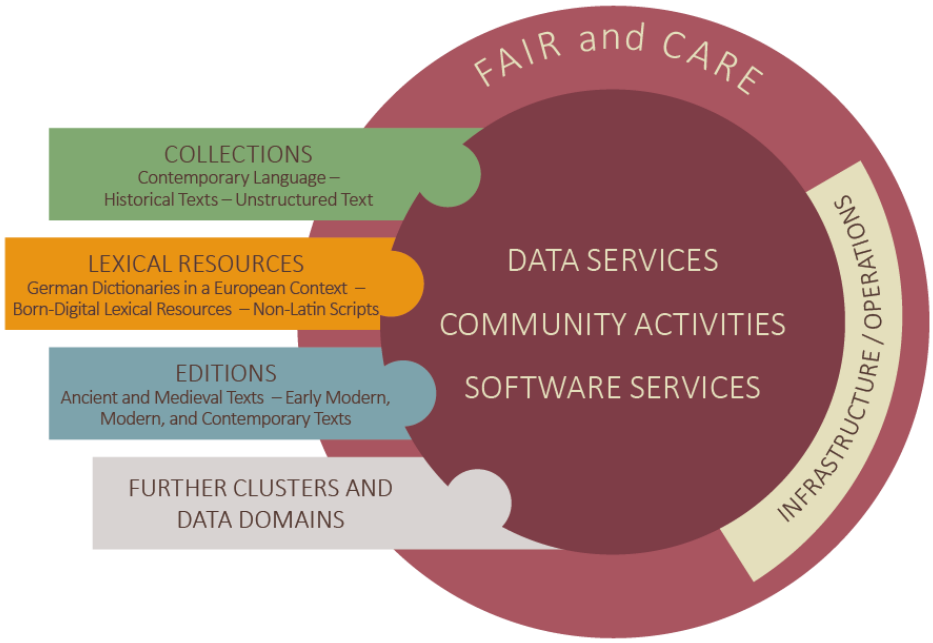

Figure 4.1 Text+ research data management approach

Text+ intends to organise its infrastructure in a distributed network of Thematic Clusters (introduced in section 3.1) that offer a systematic view of the three data domains of Collections, Lexical Resources, and Editions. The infrastructure needs to be configured in a modular and scalable fashion so that further Clusters may be added or existing ones restructured according to the communities' requests and further data domains may be addressed in Text+ in the future, so as to extend the data and service portfolio.

Services are developed and provided on two levels: As described in section 4.3.1 the Task Area Infrastructure/Operations offers generic services, which are adapted by each Cluster to the needs and research priorities of the specific communities of interest. This approach ensures efficient and consistent cross-data domain service provision while respecting the diversity of data and methods (see Objective 1: Support Methodological Diversity by High-quality Research Data. and 3, section 2). In addition, services are structured in data services, community activities, and software services. Each of these layers will be coordinated by Infrastructure/Operations across all data domains in order to build consensus on future developments and to provide common interfaces towards the NFDI as a whole.

Text+ complements the service offers brought into Text+ by the applicant, co-applicants, and participants as in-kind and leverages existing infrastructure components wherever possible to design, develop and operate its service portfolio.

Clusters will offer a rich service portfolio including, but not limited to the following services:

1. Data Services: Findability
    - Registries and registration procedures adapted to specific requirements
    - Federated metadata and content search solutions adapted to specific requirements
    - Integration of metadata and data from Cluster repositories into Text+ discovery services
    - Metadata harvesting and integration into Text+ discovery services for relevant resources hosted by infrastructure providers beyond the Text+ infrastructure
2. Data Services: Accessibility
    - Data hosting and ingestion procedures including basic identification services adapted to specific requirements
    - Long-term archiving for Cluster repositories
    - Quality assurance through repository certification and monitoring
3. Data Services: Interoperability and Re-usability
    - Metadata, Authority Data, Terminology services such as semi-automatic enrichment of data and metadata and participation in standardisation activities
    - Implementation of data and metadata conversions, including to LOD formats, and respective linking to knowledge graphs
4. Community activities
    - Data domain- and Cluster-specific consulting, training, and community involvement
5. Software services
    - Data domain- and Cluster-specific implementation of generic services and software development
    - Alignment with overall Text+ software development paradigms

All Cluster services build on generic services provided by Infrastructure/Operations:

1. Data Services: Findability
    - Registry solutions
    - Hybrid search and retrieval architecture across distributed and heterogeneous data resources
2. Data Services: Accessibility
    - Generic data repository based on quality criteria for data and metadata
    - Basic identification services including central authentication and authorisation infrastructure (AAI) and persistent identification (PID) services
    - Long-term data preservation and archiving
    - Quality assurance through certification of repositories and monitoring services
3. Data Services: Interoperability and Re-usability

- Metadata, Authority Data, Terminology services including quality management, semi-automatic enrichment processes and conveyance of specific requirements in further standard building by implementing a GND Agency for language- and text-based research data
- Metadata infrastructure based on standards and authorities, linked open data, schema management services, and tools for data integration and linkage

4. Community activities
    - Web portal as central access point to the whole service portfolio
    - Helpdesk including support research data management planning

5. Software services
    - Data Processing Pipeline
    - Software Development Support

Text+ is highly committed to support cross-NFDI service adoption, evolution and provision, based on the Berlin[125] as well as the Leipzig-Berlin[126] declaration. Wherever possible, services will be compliant with the cross-NFDI Research Data Commons and may be offered across potentially all NFDI consortia. This contribution includes, but is not limited to the following services in particular:

- Definition of standards for language- and text-based research data and metadata
- Enabling access to and analysis of language and text data
- Metadata infrastructures and federated content search
- Certification of data centres
- Generic services as mentioned above, including production ready services such as long-term preservation, persistent identification services, and AAI.

The overall mission of the RDM strategy is to significantly evolve the research data management for Text+'s communities of interest by mastering the Grand Challenges and achieving Text+'s main overall objectives (see section 2.2). All Measures are described in detail in the section 5, including milestones and deliverables as measures of success.

[125] see Glöckner et al. (2019).
[126] see Bierwirth et al. (2020).

5 Work Programme

This work programme defines a clear and concise plan of how Text+ implements its RDM strategy to achieve its objectives. The work conducted according to this programme is carried out in the five Task Areas: Collections, Lexical Resources, and Editions concentrate on the three major types of research data Text+ is focused on. The Task Area Infrastructure/Operations is responsible for the design, planning, development, and integration of the Text+ infrastructure and services. While services offered by the data domains mainly concern issues related to the modelling and representation of domain specific data, Infrastructure/Operations provides the generic services for the three data domains, which adapt these services to their specific needs and continuously evolve their requirements. The Task Area Administration implements the overall Text+ governance and coordination tasks within the NFDI.

The three data domain-related Task Areas are closely interrelated and complement each other. The main objective of the Task Area Collections is to keep stock of available text and language collections and to enrich them, for example with structural and semantic annotations and metadata. Furthermore, the Task Area Collections focuses on the integration of new collections and linking to other collections and knowledge resources including lexical resources. Lexical Resources consist of reference works of word usage: they link to collections as well as to editions by using them as source references (*Belegstellen*) and, simultaneously, are linked from collections and editions to provide enrichment and authority information. Editions are the reliable and methodological preservation, presentation, and critical commentary of all kinds of texts in various languages and writing systems. They make use of collections and authority files as a generic source for further enriching and contextualising information to support manual and intellectual analysis. The work programme of the three Task Areas is geared towards advancing each data domain individually according to Objectives 2 and 5 (see section 2.2) as well as inter-connecting them so as to promote Objectives 3 and 4.

The data domains are initially organised in eight Thematic Clusters (see their introduction in section 3.1 with Figure 3.1) bundling activities related to specific subtypes of data and research methods in one data domain. The cluster systematics is not a strictly categorial one: tasks, methods as well as contributing partners partially overlap, thus ensuring efficient communication across Clusters and the timely identification of potential synergetic collaboration. Clusters can be restructured and/or added as required or recommended by SCC decisions, which will also govern the portfolio of each Cluster.

Before presenting the organisational structure of the work programme and the approach to address data domain- and Cluster-specific tasks, we elaborate on an essential aspect for the initial work and the overall portfolio development of Text+, which are the data selection criteria.

Data Selection Criteria

The SCCs select data to be included in the Text+ infrastructure using a set of common criteria and a set of data domain- and cluster-specific criteria. The common criteria encompass the following.

Relevance: What is the relevance of the data set for a given community / communities? Relevance here means

- the proven or expected *impact*, i.e. the number of users or the number of citations in publications of a resource or its being key for a community. For instance, a resource such as DeReKo (German Reference Corpus) is central for empirical German linguistics and has a large number of users and citations, while specific digitised cuneiform tablets are important for small communities (Hethitology and Assyrology) and are a key resource for at least some of their members.
- the potential for *innovation*, i.e. how effective a data set is in addressing important (new) research questions. For example, a diachronic corpus of *Kiezdeutsch,* a variety of German spoken by some groups of young people in urban areas, may boost research on language contact and change.
- the potential for *transdisciplinarity*, i.e. to what extent a data set can reach out into other disciplines. For example, a diachronic corpus of private correspondence such as love letters is of interest for linguists, sociologists as well as historians.

Diversity: Does the data set expand the spectrum of resource types in the existing Text+ portfolio? For example, there may be extant coverage of certain text types (e.g. newspapers, magazines, novels) but little coverage of non-canonical texts ("The Big Unread" in Literary Studies) or of particular languages/varieties (low-resource settings in Linguistics). How can the representativeness and balance of the data be maintained despite the diversity?

FAIR and CARE principles: What is the data set's level of compliance with FAIR and CARE principles? Potential contributing scholars and institutions should provide a statement on the conformance with the FAIR and CARE principles.

Quality: To what extent does the data set adhere to common quality criteria? Data set providers should demonstrate that their data are complete, accurate, well-documented and adhering to community-specific standards. For example, for a reference corpus this includes standards of corpus design (representativeness, balance)[127] as well as basic text quality (for example when text was digitised).

Effort: What is the estimated effort for including a candidate data set in the Text+ portfolio? A data set may be highly relevant, but its quality may be compromised and/or FAIR and CARE principles may

[127] For a more nuanced discussion, see Kupietz (2015) and Kupietz et al. (2010).

not or only partially have been considered, which can make curation too cumbersome to be feasible. For editions, an established practice is to at least preserve the XML data if the curation of the presentation layer is too complex (proprietary software, etc.).

Data offers are continuously evaluated by the SCCs based on these criteria. At the time of submission of this application, more than 25 data contributions were registered.[128] For the incorporation of new datasets from new partners, the budget includes flex funds to support the FAIRification and CAREification of these data, which will be allocated by decision of the scientific board.

Organisational Structure

In order to ensure a shared approach on how the specific aims and objectives of Text+ are implemented, all three data domains have structured their programme in their Task Areas into five Measures as follows:

Figure 5.1 Organisational structure of Text+ based on Task Areas

Measure 1: Reference Implementation (M1)

This Measure establishes all respective workflows of a Cluster with an initial focus on selected types of data resources. The reference implementation will be based on representative data resources and services for the data domain. Data resources are provided through own funds.

[128] see https://www.text-plus.org/en/research-data/data-from-the-community/

Measure 2: Portfolio Development (M2)

Based on the reference implementation, data centres[129] will expand their portfolio regarding data and services. Clusters are obliged to integrate any data resources and services compliant with their specialisation, the technical and quality criteria set by the Text+ Scientific Board and the respective SCC. The SCC prioritises data and services to be processed based on a common and transparent set of criteria, which has to be further developed by the SCC.

Measure 3: Standardisation Activities (M3)

This Measure addresses the standardisation needs on the level of the data domain or a Cluster. Text+ partners are already involved in manifold standardisation activities on a national and international level. They will intensify their engagement from both specialist and overarching perspectives, with a particular focus on NFDI-wide aspects. This Measure includes, in particular, interoperability aspects for metadata, authority files, and terminologies as cross-cutting topics of the Humanities consortia and beyond.

Measure 4: Community Activities (M4)

Beyond the day-to-day consulting and training activities included in the Cluster portfolio, Text+ partners will offer workshops and training activities in order to address Objective 3 "Foster Transdisciplinary Co-operation" and Objective 4 "Advance innovative research". These special events include yearly workshops or tutorials on new or emerging trends in Humanities research methodologies and associated research data for language and text. The events will be offered for researchers at all stages of their careers, but will include junior researchers in particular. In addition, Text+ data centres in all three data domains will offer research data partnerships for PhD students whose dissertation projects match and would contribute to the thematic focus of one of the Text+ data centres.

Measure 5: Software Services (M5)

This Measure is subdivided into two different aspects: firstly, it includes software development tasks and services specific to the data domain. Secondly, it bundles expertise from all three data domains for software engineering and service portfolio development, coordinated by *Infrastructure/ Operations*. In this way, Text+ will foster the development of a Text+ software architecture based on common software developing paradigms across all data domains while respecting their specific requirements.

[129] see the list of centres in Text+ at https://www.text-plus.org/en/research-data/data-and-competence-centres

Funding is allocated according to the following calculation: the implementation of the data service Measures M1, M2, M3 requires a continued and sustainable service infrastructure based on institutional funding. For these Measures, a data centre (certified or committed to acquire certification) is funded with 60 person months or 35 person months, depending on the scope of the data centre. Permanent personnel will be complemented with project staff. A competence centre without technical infrastructure will be funded with 30 person months. A Cluster is typically built of at least one, often more data centres and additional competence centres. For domain-independent software tasks, each data domain is funded with 45 person months, distributed among various partners. For Measure M4 and domain-specific tasks in M5, individual calculations apply.

Infrastructure/Operations provides all necessary generic services. Its programme is structured into Measures towards Findability, Accessibility, Interoperability and Reusability. In parallel with the data domain Measures, it is also divided into Community Activities and Software Services. For all funding, individual calculation applies.

Table 5.1 Overview of Task Areas

Task Area	Measures	Responsible Co-Spokesperson(s)
Collections	Measure 1: Reference Implementation (M1) Measure 2: Portfolio Development (M2) Measure 3: Standardisation Activities (M3) Measure 4: Community Activities (M4) Measure 5: Software Services (M5)	Dr. Peter Leinen
Lexical Resources	Measure 1: Reference Implementation (M1) Measure 2: Portfolio Development (M2) Measure 3: Standardisation Activities (M3) Measure 4: Community Activities (M4) Measure 5: Software Services (M5)	PD Dr. Alexander Geyken
Editions	Measure 1: Reference Implementation (M1) Measure 2: Portfolio Development (M2) Measure 3: Standardisation Activities (M3) Measure 4: Community Activities (M4) Measure 5: Software Services (M5)	Prof. Dr. Andreas Speer
Infrastructure/ Operations	Measure 1: Findability (M1) Measure 2: Accessibility (M2) Measure 3: Interoperability and Re-usability (M3) Measure 4: Community Activities (M4) Measure 5: Software Services (M5)	Regine Stein
Administration	Measure 1: Financial Administration of the Consortium (M1) Measure 2: Administration of Flexible Funds (M2)	Prof. Dr. Erhard Hinrichs

	Measure 3: Text+ Office: Scientific Office and Infrastructure/ Operations Office (M3) Measure 4: SCC and OCC Coordination (M4) Measure 5: Co-operation within the NFDI: Cross-cutting Topics (M5) Measure 6: Co-operation within the NFDI: Sustainability (M6)	

The sections below describe the Task Areas, their objectives, their contribution to the Text+ RDM strategy, their portfolio, and their Measures in detail. The following applies for all Measures alike:

- All partners will contribute adequately to each of the Measures. Only funded partners are listed together with the Measure descriptions.
- In-kind contributions are displayed per Task Area in the funding tables section.
- Funding for staff has been allocated according to the rationale that each Measure requires both experienced staff, which ensures the continuity and quality of the services from previous work, and junior staff. Thus, the calculation includes 50% postdoctoral researchers and 50% doctoral researchers across all Measures, resulting in an average rate of 73.050 €/year for a full-time equivalent, based on the DFG personnel cost rates.[130]
- Furthermore, for each full-time equivalent, 1.800 €/year are claimed as direct project costs, comprising travel costs and other direct expenses.
- Text+ member institutions are expected to provide workspaces with appropriate technical infrastructure and alike as in-kind contributions, so no related cost categories are included in the funding request.

5.1 Collections

5.1.1 Collections and their Research Domain

We use the term Collections to refer to digital written or spoken text or language material that is kept in a digital archive, such as, for example, *Deutsches Textarchiv* (DTA), or has been compiled into a dedicated corpus, such as, for instance, the *German Reference Corpus* (DeReKo). Collections may contain material from different languages or different diachronic stages of a language and different writing systems. Collections may include material from different registers (e.g. scientific, legal, religious), text types, genres (e.g. love letters, official documents, news articles), and media (e.g. inscriptions, manuscripts, printed materials as well as digital materials, such as, e.g., e-mails, tweets, or blogs). Collections may also include audio and video recordings of speech, gestures, sign language and digital facsimiles of text-bearing objects. Materials in collections may differ in their degree of structuring: there are, for example, large collections of plaintexts for use as training corpora for some machine learning (ML) methods, while in other cases information on the internal structure, such as

[130] https://www.dfg.de/formulare/60_12/60_12_de.pdf

chapters or column layout, is available. Collections are often linked to resources in other places, for instance, dictionaries, knowledge bases or authority files.

Collections are thus at the heart of all language- and text-oriented disciplines, ranging from Literary and Cultural Studies, Linguistics and Computational Linguistics to History, Philosophy and the Social Sciences as well as small disciplines such as Tibetan and Byzantine Studies or Ancient Language Studies. Digital text/language collections of high quality are thus essential for enabling research in these disciplines, whether it is qualitative or quantitative, or uses traditional corpus-based or modern data-driven techniques. Today, text processing benefits from the use of applications based on Artificial Intelligence (AI) and Machine Learning (ML). Digital text/language collections are essential to enabling data-driven research in the Humanities and to applying text and data mining. The predominant and most successful approaches in AI are based on deep learning, a technique that requires large sets of collections.

For optimal use, collections are typically accompanied by descriptive metadata, and enriched by annotations reflecting features of the text or audio(visual) data. Such enrichment includes linguistic information, literary annotations and mark-up expressing the physical and/or logical structure of the represented object. Equally important are cross-references (e.g. as Linked Open Data (LOD)) to other documents or to knowledge bases (e.g. authority files such as GND[131] or Wikidata[132]). Annotations are derived by manual or automatic means, or a combination of both. Such enriched and structured collections lend themselves well to data harvesting and other automatic means of searching, analysing, clustering, and visualising digital content and for generating derived data, such as concordances or frequency lists.

Given the central role of collections in the text- and language-based Humanities, the Task Area Collections is intricately interwoven with the other Task Areas in Text+. For instance, the scientific texts included in the DTA may be used to build a lexical resource, such as a list of German historical scientific terms. Similarly, material from a text archive may be relevant for a specific editorial project, e.g. the first editions included in the Early English Books Online (EEBO). Vice versa, resources included in the Task Areas Lexical Resources and Editions may be linked up with resources from collections. Here, it is important to note that a specific collection may have been compiled for a specific research question. Making it usable for another purpose, e.g. for building a lexicon or feeding into an editorial project, will require some extra steps, such as, for instance, linking up authority files, generating more metadata or adding specific transcriptions or annotations (EEBO, e.g., offers only a diplomatic transcription).

[131] https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html
[132] https://www.wikidata.org

The central role of collections is also witnessed by the numerous and diverse user stories contributed to the Task Area Collections. Note again that these are authentic user stories that were collected from different communities of interest. The issues mentioned in the user stories range from corpus compilation and building including legal issues to automatic enrichment with authority files as well as annotation at various linguistic levels, e.g. named-entity recognition.

For building collections, we distinguish three kinds of user scenarios:

(a) Researchers who want to use an existing collection or combine (subsets of) different collections (commonly referred to as *virtual collections*), as described in the user stories Building Collections for Literary Scholars and Compile your own learner corpus – with Weblicht. They need federated search across repositories, possibly in combination with multilingual tools as pointed out, e.g., in the user stories Infrastructural Needs for Romance Research Data and Tourist Guides, Bazaar Histories, and the Politics of the Past in Late 20th and Early 21st Century India.

(b) Researchers who want to produce text/language collections from scratch and are looking for recommendations and best-practices or standards in research data management with guidance and active support, ideally from a very early stage of the production until its publication as re-usable research data (see, e.g., user stories Soldiers' letters of the 18th and 19th centuries or the Digital research archive on the author Hans Fallada). To archive the produced collection and to make it re-usable, they need a suitable and sustainable repository, as stated in the user story Data Archiving Support. For texts that contain sensitive information or copyright-protected material, researchers require a protected space to work with the data and to publish data subject to legal conditions (see, e.g. user stories Sharing the Visual and Textual Features of Comics and Corpus of Novels of the Spanish Silver Age).

(c) Researchers who want to use their collections for analysis but the data quality is not (yet) sufficient. For many existing digital collections, full text, viz. OCR, or metadata are missing – see the needs formulated in the user stories Full-text digitization with OCR and Facilitating research in Early Modern German texts by OCR. Here, Text+ offers consultation and active support to improve usability and FAIRness.

5.1.2 Added Value

At the end of the initial five-year funding period, Text+ will have established community-driven workflows (see section 2.2, Objective 2) for integrating extant tools and data and extended its coverage by integrating new resources as well as new communities (see section 2.2, Objective 3). Thus, Text+ Collections will offer the following added values to relevant communities – in direct response to the Grand Challenges (see section 2.1, especially **Grand Challenge II**):extant tools and data and extended its coverage by integrating new resources as well as new communities (see section 2.2, Objective 3).

Thus, Text+ Collections will offer the following added values to relevant communities – in direct response to the **Grand Challenges I** and **II** (see section 2.1):

- Text+ will act as a comprehensive support infrastructure for all issues regarding collections. It will support researchers in creating, sharing and using collections in a productive way (see section 2.1, **Grand Challenge II**).
- Text+ will provide easy and transparent processes to find and obtain data from collections freely available for research, and to deposit such data for preservation (see section 2.1, **Grand Challenge II**).
- Text+ will integrate collections currently beyond the reach of researchers in a community-driven way and at a large scale.
- Text+ will be a hub for legal and ethical expertise related to collections.
- Text+ will increase standardisation and interoperability for collections by Linked Open Data, controlled vocabularies and authority files (see section 2.1, **Grand Challenge I**, and section 2.2, Objective 5).

As a result, the improved quality, interconnectivity and interoperability as well as the sheer amount and range of data (see section 2.1, **Grand Challenge I**) will allow researchers to address novel research questions and explore novel research methods at a large scale (see section 2.1**, Grand Challenge II**, and section 2.2, Objective 4).

5.1.3 Network Collections and its Clusters

Text+ brings together ten data centres for collections in Germany, namely the IDS, the University of Tübingen, the BBAW, the DNB, the SUB, the Bavarian Speech Archive at the LMU Munich (LMU/BAS), the University of Cologne, the University of Hamburg and the Academy of Sciences and Humanities in Hamburg, as well as the CLARIN Centre for Parliamentary Data at the University of Duisburg-Essen. The initial selection of data centres was made based on the criteria of representativeness for the broader scientific community, a high diversity, quality and volume of the collections provided, as well as existing experience in the function as data centre. The combination of the expertise in the data centres from the library world and those from the academic world will facilitate the production of high-quality data (see section 2.1, **Grand Challenge I**), but also the accessibility of data across disciplines (see section 2.1, **Grand Challenge II** and section 2.2, Objective 3).

Data centres are organised in three different Clusters:

The Cluster *Contemporary Language* will be coordinated by the IDS and the University of Tübingen (UniTÜ). The data contributed in this Cluster includes speech (IDS, LMU/BAS, UniHH, AdWHH, UniK), sign language (AdWHH/UniHH), written language (IDS, UniTÜ) and other modalities. Collections are mono- or multilingual. There are special-purpose language corpora such the parliamentary minutes

(UniDUE) or language documentation of endangered and minority languages (UniK, AdWHH/UniHH). Experimental data, e.g. from psycho- and neurolinguistic experiments, will also be part of this Cluster, and will be treated as annotations or metadata. Moreover, we include derived text formats such as, e.g., lists of N-grams (see M2 Portfolio Development below).

These data are prima facie heterogeneous but demand similar approaches. Typically, a normal (orthographic) form exists to which texts can be related. Tooling for contemporary forms of written language is well developed. However, there is potential for further improvement with respect to analytical depth, e.g. using machine-learning techniques, and tools must be extended to additional, including non-European languages as work progresses. Speech processing and audio-visual data also offer specific challenges. There is considerable expertise in this cluster concerning the standardisation of resources, both for collections in general (IDS, LMU/BAS) and for metadata in particular (IDS, UniTÜ), for specific varieties (UniDUE, UniHH) and for legal and ethical issues (IDS). Regarding specialised software services, the Cluster also offers comprehensive expertise, as its members have not only experience in curating collections and working with them but have also developed open tools for important tasks in curation, pre-processing, conversion, and analysis. For instance, tools are available for annotation, speech processing or enriching text with structural and linguistic annotation, analysing it and making it searchable. Intuitive (graphical) user interfaces have been deployed as wrappers, combining and facilitating the usage of technically-oriented tools, both domestic and externally developed. This also enables reproducibility of research procedures. Data centres have contributed, e.g., to decentralised search on content and metadata in CLARIN and DARIAH, and to defining and processing virtual collections.

The Cluster *Historical Texts*, coordinated by the BBAW, deals with documents from the very early times up to ca. the first half of the 20th century. The partners demonstrate a strong focus on and expertise in creating and dealing with structured texts and documenting stages of historical German, with widely-used corpora that have already been harmonised on an annotation level (text and metadata) according to standards and best practices from the CLARIN and CLARIAH projects. Other than German-language corpora, the Cluster will also contribute and integrate data on (historical) language stages of other languages including, e.g., Spanish, English, Latin, Greek, Hebrew and Sorbian as well as non-European languages, e.g., Ethiopian, Maya, Ancient Egyptian-Coptic, and Ancient Egyptian written in hieroglyphic signs. The various languages are represented in Latin as well as non-Latin scripts.

The Cluster *Historical Texts* combines the competencies and proven expertise in the field. Federated metadata and content search facilities for text and metadata (FCS, VLO) have been implemented for selected languages and corpora, respectively, and will be made available for relevant texts not yet included. In addition to search and retrieval facilities, tools and services for analysing, visualising, annotating, comparing, and collating documents are in place and will be developed further. The

careful, criteria-based curation and ongoing integration of additional text resources will be continued grounded in the high standards set by, for example, the CLARIN-D curation projects. This includes the 'upconverting' of unstructured text documents (e.g. plain text, semi-structured OCR text from the index of digitised prints[133] and newspaper mass digitisation projects) to fully annotated research data, in co-operation with the Cluster *Unstructured Text*.

The Cluster *Unstructured Text* is jointly managed by the DNB and the SUB, which are already connected through a variety of joint activities. These include the German Digital Library (DDB) and the working group on the Collection of German Prints as well as standardisation bodies such as DINI-KIM. Most of the digitised documents in the collections of the DNB and the SUB, as is the case for most libraries and archives, are unstructured text in the sense that information on the structure of the individual text is not yet available and only implicitly contained in the data. This is due to the fact that the output of large digitisation projects usually consists of images accompanied by bibliographical and structural metadata (e.g. METS/MODS). Access to these documents has been predominately provided through digital library interfaces (i.e. viewers) that imitate analogue reading practices. As such, they are neither tailored towards machine-processing of textual data nor make use of OCR techniques. However, depending on the research objectives and methodologies, the provision of digital facsimiles is not sufficient, as computational approaches require plain text as the bare minimum.

The DNB will integrate its huge and ever-growing collection into Text+. Due to the legal mandate of the DNB, this includes all publications from publishing houses located in Germany. The DNB contributes its high expertise in legal issues. Together with the DDB, the DNB aims to incorporate the collection of historic newspapers as full text as soon as these are available. The SUB focuses on the integration of multilingual collections (over 13 million digitised pages as images) from large digitisation initiatives that provide digital facsimiles (VD17, VD18). These collections, supplemented by data from the VD16, are used for automatic text recognition in the OCR-D project. In the coming years, the results of this process as well as other resources (e.g. scientific journals of the 17th-21st century) will be integrated into and made available via the Text+ infrastructure, in compliance with the legal framework. The latter poses particular challenges and makes it necessary to provide a set of derived formats for legally protected data that enable the usage of these resources while at the same time ensuring that full documents cannot be reconstructed on their basis (see M2 Portfolio Development).

The Cluster *Unstructured Text* will be a model for the community-driven integration of collections of GLAM institutions into Text+. It will mediate between the disciplinary communities and these institutions regarding the generation and provision of digital text collections in compliance with the legal framework. The Cluster will support these stakeholders in generating unstructured texts and

[133] https://www.zvdd.de

gradually enriching it with structural information (see M4 Community Activities), resulting in collections of varying degrees of structuring.

Cluster-spanning topics. Clusters face some similar challenges. For instance, historical texts provide a wealth of spelling variation.[134] Treatment of historical spelling variants is similar to other areas of writing not conforming to standard orthography, such as computer-mediated communication (CMC) or transcripts of spoken language.[135] Similarly, legal issues with data protection or intellectual property rights and licensing concern both the Clusters *Contemporary Language* and *Unstructured Text* (see M2 Portfolio Development on derived formats).[136]

Common concerns for Text+ as a whole, such as the adaptation of tools to new languages and writing systems, also concern Collections (see M5 Software Services).

The very rich and diverse datasets offered in response to Text+'s call for data show the wealth of resources the community has to offer, but also the lack of uniformity with respect to technical details and data formats. The harmonisation, close connection, seamless integration, and/or interlinking of resources, e.g. via controlled vocabularies and authority files, will therefore be an ongoing important task in all Clusters (see section 2.1, **Grand Challenge I**). Cross-referencing multilingual text resources on the same topic with each other will be another added value, as will be a federated, multilingual search and retrieval facility (see user story Supporting Information Retrieval in and for Multilingual Scholarly Editions/Text Resources or Linguistics in non-European Languages). This topic will be addressed together with the Task Area Operations.

5.1.4 Measures

In this section, we specify the Measures to be carried out by the data centres described in section 5.1.3.

Taking into account the funding cuts, the work programme of the task area collections was adapted accordingly. The available person-months for the implementation reflect not only the cut of 29.8% of the budget, but also the fact that the individual personnel costs increase over the five years of funding and are higher than the DFG personnel cost rate of 2020, which was the basis for the budget calculation of the proposal. This means that in terms of funded person-months the reduction for some partners will be much higher than 29.8%.

All measures and most of their tasks are essential both for the objectives of the task area Collections as well as the overall Text+ goals and will be addressed in this funding period, but with limited scope and depth in comparison to the original planning. A lower degree of completeness and integration will

[134] For computational approaches to normalisation see, e.g., Jurish (2011), Bollmann (2019).
[135] see, e.g., Lusetti et al. (2018) for CMC, Schmidt (2012) for transcripts of spoken language.
[136] see, e.g., Deutsche Forschungsgemeinschaft (2015), Kupietz et al. (2018a), Arnold et al. (2020), Kamocki et al. (2020).

be achieved, only a fewer number of outreach activities including e.g. workshops, guidelines and reports will be delivered.

In particular, the funding cuts implied the following modifications of the work plan:

- Measure 1: Substantively, there is a postponement of the milestones and a focus on the ingest processes. A complete integration of all data has to be abandoned due to the cutbacks, instead, Text+ aims at an exemplary integration of prominent parts of the partners' collections.
- Measure 2: In Measure 2, the reductions again lead to a shift in activities as well as a greater focus on implementation. The report on the continuous integration of resources forms the final report.
- Measure 3: The cuts necessitated by the budget reductions also pertain to the work program in Measure 3. Under the given conditions, adoption of an ISO standard is unrealistic. The internal workshops that were originally planned for the third year will be spread out over the third and fourth years.
- Measure 4: When it comes to this important measure the cuts are particularly painful. As a consequence of the cuts, however, the number of recommendations and the number of workshops will have to be reduced.
- Measure 5: In light of the cuts, Text+ will focus on collection-specific needs assessment and infrastructure building, but the number of implementations of specific tools will have to be reduced and, if appropriate, replaced by references to already known applicable solutions.

Measure 1: Reference Implementation (M1)

Essential parts of the contributions of the centres will be integrated, and a unified ingest processes, data formats, and unified access to these resources will be set up. This will form a solid basis for integrating new collections and resources (see Measure 2: Portfolio Development). Due to the volume and diversity of the contributions, the integration will be an iterative process. With the standard-conformant preparation of text and metadata (see section 4.2), e.g. using Dublin Core, the TEI-XML-compliant DTA 'Base Format' (DTABf) as a pivot format for prints and manuscripts, or the ISO 24624:2016 for Transcription of Spoken Language, collections across institutions become truly interoperable. A conversion to DTABf with the huge contemporary written corpora of the IDS (in I5 format) is available.

When this Measure will have been completed, the following components will be in place: a first version of an open, and documented ingest processes will exist for integrating new data. In coordination with the SCC Collections, the general criteria for the selection of data to be integrated, as introduced at the beginning of section 5 above, will have been implemented with respect to Collections. Pivot formats

for all modalities (written, spoken/transcribed, audio, audio-visual, observational data) will have been selected. Unified access to the data respecting concerns of Intellectual Property Rights (IPR) and data protection and connection to central services such as the Data Management Plan Registry, Discovery Services, and Federated Content Search will have been implemented.

Funded Partners (Measures 1–3): AdWHH, BBAW, DNB, IDS, LMU, SLUni, SUB, UniDUE, UniFR, UniHH, UniK, UniTÜ, UniWÜ

Milestone	C1.1	Concept and prototypical implementation of representative resources	M18
Deliverable	C1.2	Report on the reference implementation for decentralised collections	M24
Milestone	C1.3	Exemplary integration of representative holdings of the Clusters	M36
Deliverable	C1.4	Report on the integration of the centres' holdings	M60

Measure 2: Portfolio Development (M2)

In co-operation with the SCC, the portfolio of Collections will be extended in accordance with the criteria for data selection including the FAIR and CARE principles and the criteria for data selection and workflows will be defined and refined as follows:

Ingest of new resources. The members of the SCC Collections, headed by Sandra Richter, will decide on collections and software services to be integrated into the Text+ infrastructure based on the results of the permanently open call for data from the community and according to scientific criteria. As described at the beginning of this section 5 in the description of the data selection criteria, the Clusters will provide the SCC with an estimation of the workload to include resources. Moreover, this decision includes a proposal which Cluster or partner should host the new resources. New resources will come from different stakeholders, such as individual researchers, projects or GLAM institutions as the data call submissions demonstrate. Beside digital collections, the participating GLAM institutions have large amounts of physical holdings. If required and if feasible, parts of them are digitised and integrated into collections. The SCC will also play a coordinating and steering role.

Enable re-use. To increase re-usability, resources will be provided in different formats according to the requirements of the community and the legal possibilities (see user story Contemporary European Museums Dedicated to the Rescue of Jews During the Holocaust). Data will be connected via authority files and integrated into a LOD representation (see M5 Software Services), which interconnects collections with each other but also with lexical resources, editions and beyond.

Respect IPR. Text and language data are often not free of third-party rights (Intellectual Property Rights). This holds for most texts from the 20th and 21stcenturies, because they are usually still under copyright protection, making their unrestricted distribution impossible. Despite all improvements in the legal situation in Germany, this also applies to their usage in research, e.g. for text and data mining, especially with regard to reproducibility and sustainability, see, e.g., the user story German IT-Blogs

and their Impact on the Public Discussion of Internet Policy. These restrictions are, of course, in conflict with the FAIR principles and the goals of Text+.

Derived text formats. As recently described by Schöch et. al (2020) (see also the user story Textual features derived from copyrighted texts), derived text formats, such as the Extracted Features Dataset of the Hathi Trust Library or Google N-gram Datasets, can be used to offer these protected text resources for specific purposes in a manner that does not raise legal concerns. Text+ will define and describe a set of derived text formats that maximise usability but, even when combined, can still be published and legally used. The derived formats' usefulness for research and the impossibility to reconstruct the original text from them will be systematically evaluated. In coordination with the SCC, such formats are generated on selected subcorpora., tailored to specific uses. The algorithms used will be published as part of the research process.

Protect personal data. To protect personal data, e.g. in collections of data concerning individuals' health, but also for polls and interview situations, it is necessary to ensure pseudonymisation or anonymisation to an extent that minimises the danger to rights and freedoms of data subjects while maintaining the possibility of doing research (see, e.g., user stories Building collections of social media data for religious studies and Attitudes towards regional language forms in the city). In combination with the NFDI/Text+ AAI infrastructure, this will enable research in sensitive areas.

Contribute to authority files. Authority files such as the GND, Wikidata or to some extent GermaNet provide a representation of external knowledge, as is stated in the user stories Named Entity Linking Service to enrich Textual Collections and Bibliotheca Arabica. Dealing with authority files within collections in such an elaborated way will also help to find conceptual gaps in the current state of the GND and other authority files. This Measure will document these gaps and inform the providers of the authority files in a systematic way. There will be a software tool to connect collections with authority files (see M5 Software Services, but also the service Metadata, Authority data, Terminologies of the Task Area Infrastructure/Operation).

Funded Partners: see Measure 1

Deliverable	C2.1	Annual monitoring report to the SCC	annually
Milestone	C2.2	Guidelines for integrating data into Text+/NFDI, collection-specific version	M18
Deliverable	C2.3	Report on legal aspects of derived text formats	M36
Milestone	C2.4	Integration of the first set of new resources selected by the SCC	M36
Deliverable	C2.5	Report on the continuous integration of community resources	M60
Deliverable	C2.6	-	

Measure 3: Standardisation Activities (M3)

To achieve better interoperability, Text+ will be active in the field of standardisation. Specific aspects of this in the Task Area of Collections are:

Internal standardisation activities. First, all Text+ participants and partners must, in close contact with the SCCs and the community, evaluate the current state of affairs with respect to metadata, object data, and research software regarding Collections. For example, in the realm of collections of historical prints, manuscripts, and newspapers, the DTA Base Format (DTABf) will be evaluated as an overarching encoding guideline for Text+ contributors and will be developed further in coordination with the DTABf steering committee to suit the purposes of Text+. Secondly, the integration of data from experiments will be a challenge and Text+ will have to establish interoperability between these data and the more traditional formats for collections of text and speech. Other examples for standardisation activities to be focused on are the Oxford Common File Layout and derived text formats. Once internal formats will have been selected, it will be necessary to allow input from and output to formats that are used in specific sub-communities, and also provide downstream conversion to more lightweight formats such as plain text or HTML and JSON for machine processing.

External standardisation activities. In order to meet the growing and changing requirements, standards at all levels demand an active development and close and continued interaction with professional associations in all interested Humanities disciplines via the SCC. A significant contribution to the success of Text+ will therefore be the commitment of the applicant institutions in a number of standardisation committees such as the Text Encoding Initiative (TEI), DIN/ISO, CLARIN ERIC Standards Committee and Metadata Curation Task Force, World Wide Web Consortium (W3C), the MARC Advisory Committee, BIBFRAME (Bibliographic Framework) and nestor or the Digital Preservation Coalition (DPC) for the standardisation of long-term archiving solutions, but also the networking with science funding institutions and the participation in recommendations published by these institutions.

Dissemination. Members of the community often voice the demand for advice on integrating and using resources. Standards on formats and available techniques must be provided (see, e.g., user stories Acta Pacis Westphalicae Project or Epigraphy at the BBAW). This will increase the adoption of modern data management methods, thus also increasing interoperability and facilitating re-use. Training material on these aspects must be developed and updated, access and usage must be straightforward. Moreover, a helpdesk and corresponding procedures must be implemented, and regular teaching formats will be developed.

Funded Partners: see Measure 1

Milestone	C3.1	New work item proposal for an ISO standard on derived text formats	M24
Deliverable	C3.2	Publication on relevant standards for Collections (metadata and annotations)	M18
Milestone	C3.3a	Internal workshops on standards for collections	M36
Milestone	C3.3b	Internal workshops on standards for collections	M48
Deliverable	C3.4	-	
Deliverable	C3.5	-	

Measure 4: Community Activities (M4)

Text+ Collections will offer consulting and specific workshops and courses on the following subjects:

- Data curation, including standards and research data management
- Legal issues, including derived formats
- Transfer of competencies, e.g. machine learning, including digital literacy

Curation and research data management. Collections must be proactively preserved so that they do not fall into obsolescence by a curation project's end of funding. To ensure re-usability and long-term availability of collections, research data management plans have become obligatory parts of project proposals. Funding agencies have also been developing guidelines for digitisation and data management. There is basic awareness and demand for information in the disciplinary communities. Text+ will engage in the ongoing dialogue in the area of collections on how to prepare data, as discussed in the previous Measure. It is important to transfer the results of this discussion as well as the experiences gathered during the reference implementation into the wider disciplinary communities. Concrete recommendations and points of discussions will be derived from the experiences with the reference implementation of selected data and portfolio development. By its participation in associations such as RDA, DINI and nestor, Text+ ensures that the larger context in these areas is also reflected. In co-operation with networks of expertise (e.g. FIDs and others) and funding institutions, Text+ will offer advice and suggestions to enhance the practical guidelines for digitisation by the DFG (last version 2016), the corresponding discipline-specific guidelines, the DFG Guidelines on legal issues in building corpora.

Address legal and ethical issues. As both legal and ethical issues are major challenges in preparing, using and sharing collections, it is particularly important to provide counselling and information in these areas. Contracts, intellectual property, data protection, but also the liability of service providers constrain accessibility and re-use of data. The lack of clarity in extant law and its constant evolution (e.g. the upcoming transposition of the Directive on Copyright in the Digital Single Market, or issues regarding the GDPR) particularly affect the curation and use of collections and cause a high demand for specific legal consultation as signalled, e.g., in the user story Routines for the curation/integration of research data from non-European affiliations. The volume and variety of collections complicate legal assessment. Text+ Collections will offer extensive consultation and tooling to address these issues and will contribute to the law-making process. Ethical issues in research data and collections have to be taken into account. There is a hidden bias in software due to the bias of training corpora and in the selection of text and language data for building collections. Data policies, recommendations and codes of conduct address ethical issues by raising awareness and forcing stakeholders to consider their own

bias in the decision-making process. Another example concerns questions on the appropriateness of, e.g., handing out European consent forms to people with non-European cultural background.

We will build on the profound expertise of the DNB and the IDS regarding these issues.

Transfer of competencies. As working with collections is fundamental in the Humanities, Text+ Collections will offer consulting and training for creating and working with collection data, including advice to digitise analogue text and language objects. This will include text and data mining, but also advanced digital literacy for Humanities researchers. In these contexts, the aforementioned points will also be taken up.

Funded Partners: DNB, LMU, SUB, UniTR, UniTÜ

Deliverable	C4.1	Assessment of the impact of the Directive on Copyright in the Digital Single Market and its German transposition on the legal status of Collections	M24
Deliverable	C4.2	Suggestions to the DFG on digitisation and discipline-specific guidelines	M36
Deliverable	C4.3	Guidelines on the complete research data integration process in Text+	M48
Milestone	C4.4	One workshops or courses in summer schools per year on collection-specific aspects of research data	annually
Deliverable	C4.5	Sustainability report for Community Activities	M60

Measure 5: Software Services (M5)

First, requirements specific to Collections have to be formulated for central services and the centres have to connect local collections and services to central services. Operations will provide these central services. Furthermore, there are software services which are specific to Collections.

Coupling software and Text+ resources. There is high demand by the communities not only for accessing research data, but also for easy-to-use tools for enriching the data and improving data quality in preparation for advanced analysis. According to the requirements for annotation tools, they will be linked to disciplinary context and resources, see, e.g., the user story Towards improved research results in computational linguistics via Text+. Tools for preparation and general processing as well as for conversion of text and metadata format foster the re-usability of collections as elaborated on in the user story CLiGS: Textbox. Quality assurance platforms or specific tools for collating documents exist and will be further developed. Regarding OCR and quality improvement, Text+ meets the expectations of researchers by evaluation, indication, and recommendation, as mentioned, for example, in the user story Full-text digitization with OCR. For analysis, most of the existing text and data mining tools are per se language-independent but need specific training data or adaption (see user story Science in Ancient Egypt). The standardisation of formats will allow constructing general pipelines of tool application tailored to typical workflows of data curation and research. The application to the collections in the multilingual Text+ infrastructure will also allow providing language models and processing pipelines for languages that are often underrepresented in the NLP world.

Collections will formulate the specific requirements based on the needs of the community and in close dialogue with the SCC.

Link and connect. This Measure aims at connecting the resources of the collections with authority files such as the GND and knowledge bases such as lexical resources. This is crucial for the curation of collections as demanded in user stories such as International research on bibliographical data. It will be done by an enrichment of the collection data in the form of links to the entries in the authority files. Application of named entity recognition algorithms may provide the automatic or at least semiautomatic technology to enrich the data with these links. Linking and connecting data in this way is an ongoing process due to its recursive nature and due to the arrival of ever new resources. Collection-specific software services will be provided, which make use of extant PID infrastructures. This Measure is connected to the service Metadata, Authority data, Terminologies in the Measure Interoperability and Re-usability of the Task Area Infrastructure/Operations.

Enabling innovative research. Text+ aims to provide sophisticated but easy-to-use ways to analyse text and metadata using methodologically sound, transparent and scientifically verifiable tools and services. Starting from basic term/frequency queries to more elaborated investigations of collocations, lexical and semantic features, physical and logical structures, etc., researchers will be provided with the means to examine single texts and whole collections computationally, as mentioned, e.g., in the user story HateSpeech on Twitter, and to visualise the results following state-of-the-art paradigms. Moreover, machine learning methods, originally developed for other domains in Artificial Intelligence, have high innovative potential for the Humanities as well. The Text+ data portfolio provides an ideal basis for applying such highly data-intensive methods to a wide range of research questions in the Humanities. To this end, Text+ will offer reference to solutions for the creation of customised data formats such as word embeddings and other contextualised word representations. To support this kind of research, Text+ will establish a data processing pipeline (see the Task Area Infrastructure/ Operations).

Funded Partners. DNB, LMU, SUB, UniTR, UniTÜ

Milestone	C5.1	Software Requirements for collections-specific aspects of the Federated Content Search	M12
Milestone	C5.2	Implementation of tools to connect (link) collections items with authority files (such as GND) and other knowledge bases	M36
Deliverable	C5.3	-	
Milestone	C5.4	Implementation of selected tools to exemplify the suitability of the infrastructure for new research paradigms, starting in year three	M60
Deliverable	C5.5	Sustainability report for Software Services	M60

In-kind contributions: AdWHH, BBAW, IDS, DNB, SLUni, SUB, UniDUE, UniFR, UniHH, UniK, UniTR, UniTü, UniWü.

5.2 Lexical Resources

5.2.1 Lexical Resources and their Research Domain

Words are the building blocks of human communication. Their uses within sentences, texts and multimodal arrangements are fundamental for human interaction, intellectual life, and the constitution of history and tradition. They are the fabric of social interaction, and they serve a broad spectrum of everyday communicative tasks. The usage of words is the object of a wide variety of research areas, including traditional and computational branches of (historical) lexicography, (historical) semantics, (historical) lexicology within the philologies, but also within disciplines that transcend single languages, such as language typology or comparative linguistics. Word usage is also a research object in fields such as the history of ideas, discourse studies, studies on scientific language, terminology studies, and several branches of computational linguistics and AI research.

The area Lexical Resources deals with reference works and databases of word usage. Their conception and compilation constitute a research activity that produces important results in specific disciplines, e.g. lexicography or semantics. The resulting lexical resources are used for research purposes in many areas.

The main types of lexical resources are a) dictionaries for human readers. Their scope includes different languages, different regional varieties and historical stages, different domains, and a huge variety of philological principles of organisation; b) machine-readable dictionaries; c) encyclopedias such as Wikipedia or legacy works such as Krünitz, Brockhaus or Meyers; d) terminological databases for specific knowledge areas; e) ontologies that organise lexical material according to a conceptual system or according to meaning relations; f) wordlists, e.g. frequency lists or linguistically annotated lists (morphology, syntax, semantics); g) word maps and linguistic atlases.

On the IT side of things, the research domain of lexical resources includes infrastructure units, i.e. platforms and dictionary portals where lexical resources can be accessed, searched and stored, software services for processing of lexical data as well as best practices and standards for metadata, object data and data interchange.

Lexical Resources in Text+ address four main user groups: researchers and research projects in all scientific disciplines involved in lexical research; academic and non-academic users and creators of dictionaries and word-use-related data; publishers of utility dictionaries; scientists in fields of application in which lexical resources in the broadest sense are used as components (e.g. text mining). These user groups need distinct kinds of services and support for each step of the life cycle of data production, usage and research data management.

Transition to the Digital Era and Challenges: Lexicography, as one of the oldest branches of linguistics, is a mature discipline. However, in the era of digital transformation it is fragmented[137]. Even though many dictionaries are retro-digitised, they are still not satisfactorily structured for the new digital media and do not live up to their potential. On the other hand, there is an increasing number of born-digital lexical resources including lexical reference works and word-based knowledge systems that are only loosely connected to the world of legacy dictionaries. The major challenges for the area lexical resources are unifying the heterogeneous landscape of resources and supporting the data providers in the FAIRification and CAREification process (see section 2.1, **Grand Challenge I** and **II**). More specifically, the goals of the area lexical resources are applying uniform formats to both retro-digitised dictionaries and born-digital lexical knowledge resources and making them interoperable with one another, establishing portals and generic search facilities for (groups of) lexical resources, interlinking resources to authority data, and using established standards to ensure re-use and interoperability of data.

5.2.2 Added Value

The above-mentioned challenges can best be addressed by a large and stable research data infrastructure. To this end, Text+ brings together major stakeholders in Germany as providers of infrastructure platforms for accessing, analysing, and storing data, and as providers of lexical data or software services.

The following goals are a response to the challenges mentioned above. They are matched by 25 user stories on lexical resources (see Text+ website;[138] the most illustrative are cited below). The focus on user stories is in accordance with Objective (see section 2.2).

At the end of the five-year funding period, Text+ will have achieved three general (1-3) and three cluster-specific objectives (4-6) in the data domain Lexical Resources that will result in major added values with respect to the current situation.

1. A common search interface to lexical resources will be offered for each of the Thematic Clusters (see user story Etymological Dictionary of Medieval German, see also section 2.2, Objective 5).
2. The resources of all participating data centres will be integrated in the LLOD cloud[139], and more specifically, semantically linked to common data hubs such as Wikidata, Wikipedia, and Wiktionary, as well as GND, WordNet-RDF, OntoLex-Lemon (see user story Discourse Analysis).[140]

[137] see https://elex.is/objectives/
[138] https://www.text-plus.org/forschungsdaten/user-stories/
[139] https://linguistic-lod.org/llod-cloud
[140] see Tittel/Chiarcos (2018).

3. Community-driven workflows will be established to integrate new lexical resources and services into the Text+ infrastructure (represented by the data centres described below) according to established standards and best practices (see user stories Dialect Dictionaries, and Word family database of historical German; see also section 2.2, Objective 2).
4. A particular focus will be set on the comprehensive coverage of German dictionary resources in all their varieties and historical stages. The goal is to create a one-stop infrastructure that makes major academic dictionary resources findable and accessible with one single query. A further added value is to connect dictionary resources with resources from Collections and with the results of scholarly research in word usage and word history (see the following user stories: (1) Word Bibliography Database of the Göttingen Academy, (2) Old High German Dictionary: Supplements, (3) Old High German Dictionary: External Links, and (4) Digital vocabularies and XPath-Searches on the Web; see also section 2.2, Objective 1).
5. Born-digital lexical knowledge resources contribute significantly to the development of semantic language models that are used in a variety of fields, including computational literary studies, history and psycholinguistics as well as in life sciences and Artificial Intelligence (see also section 2.2, Objective 3). In all these fields lexical resources enrich existing language models with a semantic dimension (see user story Topic Detection and Cluster Labeling with Dornseiff), including models for determining semantic similarity, a key concept in technologies such as semantic information retrieval. Terminologies benefit from this connection since the lexical-semantic resources provide a semantic backbone to the terminological data and a bridge to non-terminological uses of a particular term.
6. Right from the outset, the problem of non-Latin scripts as well as under-resourced languages will be addressed and reference implementations for the ingest of new resources will be provided (see section 2.2, Objectives 3 and 4). This includes languages with a current lack of adequate digital data (see user story Integration of lexical data).

5.2.3 Network Lexical Resources and its Clusters

Text+ brings together seven data centres for lexical, lexical-semantic, ontological and similar resources in Germany, namely a) the *Zentrum für Digitale Lexikographie der deutschen Sprache* at the BBAW (BBAW-ZDL), b) the Leibniz Institute for the German Language, Department of Lexical Studies (IDS-Lexik), c) the Trier Centre for Digital Humanities (UniTR-WbNetz), d) the University of Tübingen Data and Competence Centre (UniTü), e) Saxon Academy of Sciences and Humanities, Leipzig, (SAW), f) Data Centre for the Humanities in Cologne (DCH) and g) Research Centre for Primary Sources of the Ancient World at the BBAW (*Zentrum Grundlagenforschung Alte Welt*). Detailed descriptions of the data centres are available via the Text+ web site.

The criteria for the initial selection of data centres were the following: sufficient technical maturity, representative user groups addressed within different communities, a convincing added value proposal as well as sustainable in-kind contributions for backing the offer. In the course of Text+ and according to the decisions taken in the SCC, other data centres will follow.

The heterogeneous landscape of lexical resources is best matched by grouping the selected data centres into three Thematic Clusters that address a broad range of resources for contemporary and diachronic perspectives on language and text: A) *German Dictionaries in a European Context*, B) *Born-Digital Lexical Resources*, and C) *Non-Latin Scripts*.

The cluster structure seeks to strike a balance between the major task of being the world leader in the digital lexicography of German in all its historical stages, regional diversifications and specialist terminologies, and the goal of being well-embedded in, and contributing to, the international state of the art in producing, curating and supporting the use of digital lexical resources of all types. All Clusters closely collaborate with international partners, European as well as further afield, and conceive of working on and with lexical resources as a dynamic and highly interactive field.

The Cluster *German Dictionaries in a European Context* brings together four major providers of dictionary resources for German: they host, curate, compile and publish central dictionary and legacy encyclopedic resources in renowned and widely used portals, including the online German vocabulary information system DWDS[141] (BBAW-ZDL), OWID (Online German Vocabulary Information System, *Online-Wortschatz-Informationssystem Deutsch*)[142], OWID^{plus} and the loanword portal (IDS-Lexik), the *Wörterbuchnetz* at UniTR-WbNetz[143], and the *Wortschatzportal* at the University of Leipzig / SAW with the thesaurus project *Dornseiff.* The Cluster sets its focus on retro-digitised dictionaries and on lexical resources addressing primarily human users. These comprise general language contemporary and historical dictionaries (e.g. *DWDS*, *Deutsches Wörterbuch* by Jacob and Wilhelm Grimm), dialect dictionaries (e.g. *Pfälzisches Wörterbuch*) and dictionaries of language for special purposes (e.g. *Wörterbuch der Winzersprache*, *Wörterbuch des Protestdiskurses* 1967/1968, *Deutscher Wortschatz nach Sachgruppen* by Dornseiff).

The participating institutions have a strong background in metadata, standard formats and APIs for search interfaces. Moreover, the Cluster reaches out to lexical resources of other European languages during the course of the project. The partners are involved in the European projects CLARIN and DARIAH and the European ELEXIS[144] network.

The Cluster *Born Digital Lexical Resources* brings together major stakeholders with a broad coverage of European and non-European languages, namely UniTü, SAW, DCH. The participants host and curate

[141] see https://www.dwds.de/ and https://www.zdl.org/ and Klein/Geyken (2010).
[142] see Müller-Spitzer (2010).
[143] see http://www.woerterbuchnetz.de/cgi-bin/WBNetz/setupStartSeite.tcl and Hildenbrandt/Moulin (2012).
[144] see https://elex.is/

resources that are used by academic audiences, industrial and commercial enterprises and the general public. With *GermaNet*[145], UniTü provides one of the largest word sense databases available, which is connected to WordNets of more than fifty languages. Through its project *Leipzig Corpora Collection,* the SAW provides lexical resources for more than 250 languages, covering a large variety of those languages that are represented in the World Wide Web with a focus on languages with insufficient digital lexical resources. The DCH has a special and internationally recognised expertise in multimodal lexical databases as they are nowadays typically compiled in field work and often also by grassroot initiatives in small communities. All participants of this cluster are active in the LLOD cloud, an important foundation for cross-disciplinary integration tasks and applications.

The third cluster is the Cluster *Non-Latin Scripts and Under-resourced Languages* (referred to as Cluster *Non-Latin Scripts*). Languages represented in non-Latin scripts such as Arabic, Persian, Sanskrit, Japanese, Thai, etc. typically pose special challenges for the creation and maintenance of lexical resources. This is due not only to the special characters used in these scripts. More importantly, these languages typically pose special problems regarding lemmatisation (What are the head units?), the organisation of lexical entries (Which morphological derivations are included in an entry? How are these ordered?), the proper statement of meanings and meaning relations, the implementation of search algorithms and many others. While non-Latin scripts are typically linked with such problems, similar problems often also arise in lesser resourced languages written in a Latin script, which are therefore also included in this cluster.

Data centres participating in Text+ which have a special expertise in this area include the DCH that has a longstanding expertise in lexical resources written in Indic scripts, including the most-used family of Sanskrit dictionaries available on the web. Similarly, the BBAW maintains the world-leading digital lexical database for Ancient Egyptian, the *Thesaurus Linguae Aegyptiae (TLA)*. While these resources are based on earlier printed works, the SAW works on the automatic compilation of lexical resources on the basis of web-based corpora for more than 250 languages, including Chinese, Georgian, Korean, etc.

The focus of this cluster is on improving digital methods for creating lexical resources in non-Latin scripts and for other types of lesser resourced languages which have the same functionality as the resources available for better known languages, such as full searchability, automatic linking to textual sources and to LLOD, different extraction options, and so on.

[145] see Hamp/Feldweg (1997) and Henrich/Hinrichs (2010).

5.2.4 Measures

Taking into account the funding cuts, the work programme was adapted accordingly. The available person-months for the implementation reflect not only the cut of 29,8% of the budget, but also the fact that the individual personnel costs increase over the five years of funding and are higher than the DFG personnel cost rate of 2020, which was the basis for the budget calculation of the proposal. This means that in terms of funded person-months the reduction for some partners can account for up to 40%.

All measures and most of their tasks are essential for the Text+ objectives and will be addressed in this funding period, but with limited scope and depth in comparison to the original planning. A lower degree of completeness and integration will be achieved, only a fewer number of outreach activities including e.g. workshops, guidelines and reports will be delivered.

In particular, the funding cuts implied the following modifications of the work plan:

- Measure 1: Later implementation of Deliverables LR 1.1 and LR1.2; LR1.3 reduced to selected contributions to be integrated
- Measure 2: LR 2.1 is a contribution to the report rather than an independent report, LR 2.2 is left out, LR 2.3 and LR2.4 are completed later, LR2.6 is reduced to base functionality
- Measure 3 is reduced to Milestone LR 3.3 and Deliverable LR 3.4
- Measure 4: LR 4.2 have a reduced number of courses and workshops, LR 4.3 is left out
- Measure 5: LR5.1 is completed later, as is LR 5.4; LR5.5 is reduced to selected data

Measure 1: Reference Implementation

With the start of Text+, all data centres will provide and operate their lexical resources, services and web-platforms. Regarding the Cluster German Dictionaries in a European Context, where a close co-operation between the data centres is necessary to achieve a strong degree of interoperability, there is an agreement of BBAW-ZDL, IDS-Lexik and UniTR-WbNetz to co-operate on common data standards for their respective resources and on the establishment of overarching search facilities for all data. All data centres will prepare user-oriented guidelines in order to facilitate the ingest of new resources. By the standard-conformant (see M3 Standardisation Activities) preparation of text and metadata the resources become truly interoperable and will be integrated into a common environment (see M5 Software Services).

Funded Partners (Measures 1–3): BBAW, IDS, SAW, UniK, UniTR, UniTÜ

Deliverable	LR1.1	User-oriented guidelines for data ingest	M18
Deliverable	LR1.2	Concept for the implementation of decentralised dictionary platforms	M36
Milestone	LR1.3	A selection of in-house contributions from each partner integrated	M60

		into common environment	

Measure 2: Portfolio Development (M2)

The Cluster Born-Digital Lexical Resources works on the continuous development of its resources portfolio with a focus on extending lexical coverage, simplifying the referencing, and accessing its resources in the context of the FCS and the LLOD cloud. Access and findability of their resources will be extended to new domains and use cases. The improved portfolio will specifically help connecting lexical entries with resources from Collections and with external knowledge graphs (incl. the NFDI knowledge graph146) as well as text analysis and integration tasks in multilingual environments. Another focus will be set on improving the availability of lexical resources for non-Western and lesser-resourced languages.

Required Development: (1) Extension of GermaNet by the linking of senses to Wiktionary sense descriptions and Wikipedia entries and the coupling of GermaNet lexical units with statistical word representations such as "word embeddings" (UniTü)147, (2) continued curation of contemporary web-based lexical resources with a focus on lesser-resourced languages (SAW), and (3) improvements of the connectivity of available resources with a focus on using established LLOD formats (SAW, UniTü). The latter also includes participation in the FCS standardisation process for lexical resources which is guided by the German Dictionaries in a European Context Cluster.

Funded Partners: see Measure 1

Deliverable	LR2.1	Contribution to the annual monitoring report to the SCC	annually
Deliverable	LR2.2	-	
Deliverable	LR2.3	Overview of writing systems and character encoding standards and article formats used in the domain	M24
Milestone	LR2.4	Implementation of APIs for the access of distributed lexical data	M48
Milestone	LR2.5	Integration of new lexical resources and enhancement of existing data	M48
Milestone	LR2.6	Common environment with base functionality for the access of lexical resources	M60
Deliverable	LR2.7	Sustainability report for Portfolio Development	M60

Measure 3: Standardisation Activities (M3)

The work in this Measure focuses on the re-use as well as the active development of common standardisation activities related to lexical data, including general TEI, TEI-Lex0 format as a best practice format for an interoperable encoding of retro-digitised dictionaries developed in the ELEXIS project, LMF (Lexical Mark-up Framework, ISO-24613:2008) for encoding machine-readable

146 see Schimmler (2020).

147 see Dima/Hinrichs (2015).

dictionaries; the W3C standard for ontological data (OWL) and the related data model, ONTOLEX, and the lexicon model for ontologies, LEMON, and finally LLOD best practices and data models in the field of sharable machine-readable lexical resources.

This Measure addresses all providers and users of lexical or ontological data who want to use standards and best practices to provide FAIR data. In addition, centres maintain expertise in the field (e.g. as active members of the standardisation bodies and communities) and raise awareness for standard-conformant data. All involved partners will contribute to link their lexical data with GND subject headings in co-operation with Text+ GND agency (see section 5).

Specific contributions to the development of standards are provided by BBAW-ZDL. As one of the founding members of TEI-Lex0, the BBAW will actively accompany its ongoing development.

Funded Partners: see Measure 1

Deliverable	LR3.1	-	
Milestone	LR3.2	-	
Milestone	LR3.3	A selection of Lexical Resources available in standard-conformant format (cf. LR 1.3)	M60
Deliverable	LR3.4	Sustainability report for Standardisation Activities	M60

Measure 4: Community Activities (M4)

Community activities in the field of lexical resources address two target groups: data providers and (academic) users of the data. In general, the following communities are addressed by these Measures: the Philologies of modern, historical and ancient languages, as well as Computational Linguistics and neighbouring disciplines (see section 2 on the scope of the project).

Three categories of Measures are planned: consulting, teaching and dissemination.

Regarding consulting, the data centres in Text+ will help projects and individual researchers, including PhD students) to develop research data management plans (all centres). They will provide advice in consulting in the field of standard conformant digitisation and data curation of legacy dictionaries (BBAW-ZDL, UniTR-WbNetz); in the field of evaluation, i.e. on how the use of resources can be examined empirically (IDS), and consulting on how to improve the accuracy of technologies such as semantic information retrieval, sentiment detection or automatic reasoning for lexical-semantic resources, addressed to both academic and commercial users (UniTü, SAW).

Regarding teaching and dissemination, all centres will provide in-house teaching and small-scale workshops (see the introduction of section 5 for more details on the thematic orientation of these activities). Teaching material will be made publically available after the workshops. Moreover, Text+ will foster the integration of its expertise in standards and into the specific curricula (e.g. Philologies, European Master in Lexicography). All centres are involved.

Funded Partners: BBAW, UniK

Deliverable	LR4.1	Concept for consultation and teaching activities	M12
Milestone	LR4.2	Inhouse courses and workshops	M24, M48
Deliverable	LR4.3	-	
Deliverable	LR4.4	Sustainability report for Community Activities	M60

Measure 5: Software Services (M5)

All centres of the lexical resources research area are involved in the construction, application and maintenance of services. They co-operate closely with the Task Area Infrastructure/Operations (see section 5.4 for further details), The centres provide data or resource specific software tools and services while the Task Area Infrastructure/Operations focuses on the integration into central applications or research data independent frameworks. We will focus on two general tasks that are central for all data centres (Federated Content Search and Linked Open Data) as well as on one cross-cutting task (Cascaded Analysis Broker) Additional specific services will be added in the course of the Text+ consortium according to the decisions taken by the SCC.

Federated Content Search (FCS): FCS is a standard that allows accessing distributed linguistic resources via a central search engine. Resources are made available via local FCS endpoints which are implemented and run by every participating data centre148. Every FCS endpoint acts as a bridge between the data model and query engine used at a specific data centre and the standardised data model and transport protocol specified by the FCS. The responsibility of the participating data centres is to curate these data in such a way that they are interoperable and can be integrated into the FCS infrastructure via a standard-compliant FCS endpoint (see M2 Portfolio Development). The lexical resources centres are responsible for handling their local endpoints and the extension of the current FCS specification for lexical data models and suitable query mechanisms. The responsibility of the central operations unit is to provide the general framework and the search facilities. The latter includes the implementation of a web-based aggregator portal specifically adapted for querying lexical resources via the adapted FCS specification and a suitable representation of the aggregated results.

Linked Open Data. A major added value for Lexical Resources is to make the existing portfolio accessible to the overall LLOD Data cloud and to cross-domain knowledge graphs in the context of large-scale research infrastructures. All data centres strive for applying these vocabularies to those data sets that they curate and express requests to include detailed provenance data, data quality statements, statements about the legal terms of data re-use and similar information types.

The division of labour between central operations and the data centres that are part of this task is as follows: the central operations supports the implementation of necessary adaptations of relevant lexical standards (such as OntoLex/Lemon, WordNet-RDF etc.) to the needs of the data centres and the transformation of existing data into these formats. It also provides the Measures for the

[148] see Stehouwer et al. (2012).

integration of these data sets into larger structures, including the larger NFDI Text+ knowledge graph and more generally the **Linked Open Data Cloud.** The latter might include, for some centres, the aggregation and harvesting of the relevant data.

Funded Partners (domain-specific work): BBAW, SAW, UniTÜ

Funded Partners (cross-domain work): BBAW, GWDG, UniTÜ

Deliverable	LR5.1	User driven Requirement-Analysis for FCS and LOD	M18
Deliverable	LR5.2	Specification of services	M24
Milestone	LR5.3	Adaptation of CAB-models for OCR (18th/19th c.)	M36
Milestone	LR5.4	Services and converters for FCS	M60
Milestone	LR5.5	Selected data of each partner available in LOD format, accessible via portal and FCS	M60
Milestone	LR5.6	Adaptation of CAB-models for OCR (17th c.)	M60
Deliverable	LR5.7	Sustainability report for Software Services	M60

In-kind contributions: BBAW, SAW, UniK, UniTR, UniTü.

5.3 Editions

5.3.1 Editions and their Research Domains

Scholarly editions are the critical representation of historical documents[149] and essential for text- and language-based research in many fields of the Humanities and beyond, both as a prerequisite for research and as impetus for future work. Editions consist of the reliable and methodological preservation, presentation of, and critical commentary on all kinds of texts in various languages and writing systems, but also of other media such as images, music, or film. The methodologies of textual scholarship (*Editionswissenschaft*) allow for accessing and mastering the diversity of the Humanities' sources and data. They ensure that scholars document, annotate, and represent texts as reliable sources for research and teaching, often including the authorisation and canonisation of readings. The fundamental diversity of sources and data, languages and writing systems as well as discipline-specific research perspectives and outputs are reflected accordingly in Text+ to master the complexities of **Grand Challenge I** (see section 2.1).

A multitude of editorial models have evolved based on disciplines (e.g. philology, history, philosophy), types of source material (e.g. manuscripts, printed books), and types of edited content (e.g. letters, diaries, notebooks, scholarly texts and commentaries, charters).[150] The most generic of these editorial models are: (i) *documentary* (or *diplomatic*) *editions*[151] focusing on material aspects of the text, covering language variation and material detail; (ii) *genetic editions*[152] studying the writing process and thus establishing different stages of a text; (iii) *historical-critical editions*[153] focusing on different versions and variants of a text, and examining its creation and transmission. All these editorial models

[149] see Sahle (2016).
[150] https://www.text-plus.org/en/research-data/editions-list-en/
[151] see among others Pierazzo (2014); Kline/Holbrook Perdue (2008).
[152] see among others Hulle (2016); Brüning et al. (2013).
[153] see among others Damon (2016).

can be accompanied by a critical apparatus, intertextual references, indices, and commentaries, which constitute research data. Furthermore, the encoding and representation of texts in non-Latin scripts present a special challenge. The creation of scholarly editions is and will be a continuous task, both because merely a small part of the entire surviving cultural heritage has been edited so far, and due to the necessity of integrating new sources, methodological advances, and functions into the production of new or revised editions. Currently, there are approx. 1,500 mentions of "Edition" in the DFG database GEPRIS[154] and 111 projects funded by the Union of the German Academies are tagged as 'editions' in the database AGATE[155].

Since editions retain their value over time (e.g. when no newer edition is at hand or when they reflect specific editing principles, see user story Scholarly Editing), a common digital strategy must include existing projects, some of which may date back several decades. This concerns not least the handling of media change, such as the coexistence and interplay of analogue and digital editions. Furthermore, digital editions often rely on in-house developments; thus, their technical solutions are difficult to re-use and their data hardly interoperable. This results in different requirements and speeds for research data management.

The creation, provision, and use of *digital* scholarly editions constitute a new research paradigm, advancing innovative research (see Objective 4). In general, the 'openness' of the digital medium furthers the edition of texts that do not belong to the established canon; as such they question the traditionally person-centred Literary Studies and Historiography (see user stories Authority files in digital scholarly editions of correspondence and A standardized metadata format for linking the Leibniz Edition). Furthermore, digital editions offer significantly more possibilities compared to printed editions, e.g. for documenting text witnesses and contextual materials, for interlinking them with one another (e.g. correspondence networks, see user story Distributed editing of 19th century correspondences) or with other types of texts and resources (see user story Literary editions using the example of "Poetry of the German Middle Ages"). They also allow for different views on the text (e.g., "diplomatic" and "normalised" text renderings, see user story Literary editions using the example of "Poetry of the German Middle Ages"). Since the presentational layer significantly contributes to the epistemic value of the data both creators and users have a vital interest in the sustainability of all components of digital editions (see user story Scholarly Editing). Eventually, data and metadata of critical editions can be integrated into new corpora and be further analysed with digital methods such as topic modelling, stylometry, and network visualisation (see user stories On the benefits of

[154] GEPRIS (German Project Information System) is an online database that provides information on current research projects funded by the German Research foundation: https://gepris.dfg.de/gepris/OCTOPUS?language=en.

[155] AGATE is a current research information system (CRIS) for the European Academies' research, initially covering Germany and Switzerland and is being continuously expanded (A European Science Academies Gateway for the Humanities and Social Sciences): https://agate.academy/

information infrastructures for small Humanities disciplines and Graph Models for the Genesis of Goethe's Faust).

5.3.2 Added Values

The diversity of players. The complex field of digital editing concerns mainly four groups of actors: (1) creators of editions (e.g. Humanities scholars, research software engineers, information scientists), (2) users of editions' presentation layer and/or pertaining research data (i.e. re-use) (3) institutions curating and/or operating these editions on a long-term basis (memory and research institutions, data centres, research infrastructures), and (4) economic actors (publishers, software companies).

Different speeds and negotiation processes. The key objectives of the Task Area Editions are to convey digital editorial models to the disciplines, and to foster re-use, best practices, and transdisciplinary synergies. The differences in the project-specific requirements (e.g. already edited materials in PDF-format, see user story A standardized metadata format for the Leibniz Edition), the state of the technologies used, as well as the degree of acceptance of digital representation forms bring about different starting points and speeds in adopting digital methods, which must be taken into account when building a research-driven infrastructure. Since the availability and interoperability of edited texts is a decisive added value for many disciplines (see section 2.2, specifically Objective 1), the relationship between sufficient (i.e. sensible and feasible) standardisation and the necessary openness towards innovation must and will be constantly negotiated within the Text+ community.

Re-use and innovative impulses. Text+ is determined to improve the availability of reliable texts and to unlock the enormous potential included in the adoption and development of new methodologies and techniques in line with **Grand Challenge II** (see section 2.1). A cultural change by all players in the field of digital editing is required to enable scholars to engage in innovative use of research data, e.g. based on machine learning approaches (see section 2.2, specifically Objective 1 and Objective 4).

Sustainability. At their core, digital editions are research software systems which have to be curated accordingly. From this perspective, it is helpful to distinguish between their data layers (e.g. images, full text, marked up text, multimodal annotations and interlinking, schemata, authority data) and presentation layers (source code, GUIs, APIs, web applications). Software inevitably decays and without maintenance quickly loses its functionality, until it eventually disappears completely, and in the worst case takes the data with it. While the problem of archiving the data layer is technically solved by the provision of certified repositories and data centres, there is a massive need of services to curate the presentation layers (see user stories Digital Humanities and Medieval Literary Studies, Literary editions using the example of "Poetry of the German Middle Ages" and Scholarly Editing). As software systems, digital editions are often running without backup of institutional research data management solutions including a long-term operational concept. Further, the current funding conditions do not include operating and curation costs beyond the project duration. The development of strategies and

best practices for the preservation and maintenance of the presentation layers of editions and thus of their epistemic value provides a decisive added value for a crucial challenge that is still completely unsolved.

Scientific community. The Task Area Editions offers a dynamic structure that follows a bottom-up approach in order to enable as many members of the communities as possible to participate. Thus, the emphasis of its layered portfolio lies on Community Activities, addressing **Grand Challenge II**, with a focus on assistance, consulting, and training, as well as on fostering the use of standards for data and software curation and quality assurance. Starting with two initial Clusters (see section 5.3.3), the Task Area Editions will be continuously developed along the lines of the specific needs of the communities. In line with **Grand Challenge III** (see section 2.1), this is the task of the SCC via (i) the integration of representatives of the disciplines, either as participants (SCC members, specific Measures) or supporters,[156] (ii) the integration of professional associations and research institutes,[157] (iii) the description and the development of model solutions concerning formats, presentation, and software for digital editions, and (iv) in co-operation with international partners: e.g. with the national initiatives in Switzerland (NIE/INE)[158] and Austria (KONDE)[159], in ALLEA, CLARIN-EU, DARIAH-EU, with the *Union Académique Internationale* (UAI), as well as with individual international partners, e.g. the KNAW Humanities Cluster (HuC)[160], and the Institute for Textual Scholarship and Electronic Editing (ITSEE)[161]. The SCC will be adapted and supplemented by other representatives named by the professional associations in the course of the NFDI development. Since editions are foundational in many (if not all) disciplines in the Humanities, they represent an ideal object of interdisciplinary communication in line with Objective 3 Transdisciplinary co-operation (see section 2.2). This also opens up interfaces to other consortia from the Humanities and cultural studies.[162]

5.3.3 Network Editions and its Clusters

The institutions participating as partners in the Task Area Editions have many years of extensive and broad experience with digital editions and their life cycle.[163] Within the network Editions, there are two Clusters structured along common Western European historical periodisations, which reflect the change from a manuscript culture to a print culture, partially accompanied by methodological differences in edition philology. However, since there are many similarities regarding the creation,

[156] see https://www.text-plus.org/en/about-us/participating-institutions/, https://www.text-plus.org/en/about-us/academic-societies/ and https://www.text-plus.org/en/about-us/further-partners/
[157] https://www.text-plus.org/en/about-us/academic-societies/
[158] National Infrastructure for Editions, https://www.nie-ine.ch/
[159] http://www.digitale-edition.at/
[160] https://huc.knaw.nl
[161] https://www.birmingham.ac.uk/research/itsee/index.aspx
[162] see Brünger-Weilandt et al. (2020).
[163] https://www.text-plus.org/en/research-data/data-and-competence-centres/

standardisation and maintenance of data and software, the two Clusters operate in close co-operation and share tasks as needed.

The Cluster *Ancient and Medieval Texts* is coordinated by the North Rhine-Westphalian Academy of Sciences, Humanities and the Arts (NRWAW). Contributing partners are the Darmstadt co-operation (DACo), the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), the Academy of Sciences and Literature Mainz (AdWMZ), the Herzog August Library Wolfenbüttel (HAB), and the Salomon Ludwig Steinheim Institute of German-Jewish Studies Essen (STI).

The Cluster *Early Modern, Modern, and Contemporary Texts* is coordinated by the Göttingen State and University Library (SUB). Contributing partners are the BBAW, the NRWAW, the HAB, the DACo, the German National Academy of Sciences Leopoldina, the Max Weber Foundation (MWS), and the University of Paderborn (UniPB).

The editions provided by these institutions cover all types of editions (see above) and the majority of Humanities disciplines: Ancient Cultures (101-02 Classical Philology[164], 101-04 Ancient History[165]); History (102, all sub-disciplines[166]); Art History, Music, Theatre and Media Studies (103-01 Art History[167], 103-02 Musicology[168]); Linguistics (104-03 Historical Linguistics[169]); Literary Studies (105, all sub-disciplines[170]); Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies (106-01 Social and Cultural Anthropology and Ethnology[171]; 106-04 Islamic Studies, Arabian Studies, Semitic Studies[172]; 106-05 Religious Studies and Jewish Studies[173]); Theology (107, all sub-disciplines[174]); Philosophy (108, all sub-disciplines[175]); Educational Research (109-01 General Education and History of Education[176]). The editions cover all historical periods and comprise European as well as non-European cultural heritage. As such, they deal not only with Latin and other alphabetical scripts, but also non-Latin and non-phonographic scripts and sign systems.

In the following, the individual areas of competence, common to partners of both Clusters, to holistically support the creation, operation, and usage of digital editions are outlined in detail.

---

[164] The following footnotes list some exemplary projects. https://papyri.uni-koeln.de/ [NRWAW]; http://iq.bbaw.de/ [BBAW]
[165] http://kfhist.awk.nrw.de/ [NRWAW]
[166] https://haeckel-briefwechsel-projekt.uni-jena.de/ [Leopoldina]; https://edition-humboldt.de/ [BBAW]
[167] https://hainhofer.hab.de/ [HAB]; https://architrave.eu/ [SUB]
[168] https://beethovens-werkstatt.de/ [UniPB]
[169] https://wvh-briefe.bbaw.de/ [BBAW]
[170] e.g. https://fontane-nb.dariah.eu/ [SUB]; https://www.jeanpaul-edition.de/ [BBAW]
[171] e.g. https://mayawoerterbuch.de/ [NRWAW, SUB]
[172] e.g. https://corpuscoranicum.de/ [BBAW]
[173] e.g. http://www.steinheim-institut.de/cgi-bin/epidat?release=beta [STI]; https://www.sub.uni-goettingen.de/projekte-forschung/projektdetails/projekt/maps-of-god/ [SUB]
[174] e.g. https://bdn-edition.de/ [SUB]: http://ntvmr.uni-muenster.de/ecm [NRWAW]; http://www.adwmainz.de/projekte/das-buch-der-briefe-der-hildegard-von-bingen-genese-struktur-komposition/projektbeschreibung.html [AdDWMZ, DACo], https://schleiermacher-digital.de/ [BBAW]
[175] e.g. https://averroes.uni-koeln.de/ [NRWAW], https://schleiermacher-digital.de/ [BBAW]
[176] e.g. https://www.sub.uni-goettingen.de/projekte-forschung/projektdetails/projekt/klaus- mollenhauer-gesamtausgabe/ [SUB]

Data modelling. The partners have extensive expertise in creating editorial models based on the TEI guidelines as well as on other models (graph data models, relational and non-relational models) for all types of editions and a wide variety of source materials and text carriers. This includes non-Latin scripts in terms of data modelling, mark-up, and representation, e.g. Classical Arabic or Modern Standard Arabic (NRWAW, BBAW, SUB), Hebrew or Rabbinic Hebrew (STI, NRWAW, SUB), Ancient Greek (BBAW, NRWAW), Aramaic (SUB), or Classic Maya (NRWAW, SUB).

Using authority files and Linked Open Data. The usage of as well as the contribution to authority data (especially GND) is fundamental in the work of the partners, and as such the interlinkage of the latter with external resources. Furthermore, the partners actively contribute to the creation of new, high-quality authority files or their enrichment by providing information that has been gained through editorial processes. With regard to LOD, the partners have experience with using RDF data to collect contextual information about the individual editions and their contents.

Development and provision of digital editing tools. The partners have profound experience in developing, provisioning, and maintaining tools for digital editions. This covers a whole range of solutions: smaller, specialised scripts and applications for data conversions (e.g. *XTriples*[177]) or API-driven services (e.g. *correspSearch*[178]), publication frameworks (e.g. *SADE*[179]), specialised environments for multilingual editions (e.g. *DARE*[180]), graph-based annotation tools (e.g. *SPEEDy*[181]), and full-fledged editing environments (e.g. *ediarum*[182], *TextGrid*[183]). In addition, the partners are also very familiar with the majority of digital editing tools provisioned by third parties, both scientific and commercial.

Publication of digital editions as web portals and/or as print publications. The partners have extensive experience in publishing digital editions as web applications, especially using XML databases with corresponding X-technologies for rules, data storage, data access, data conversion, and data output. Additionally, the partners also employ other, non-TEI-centric approaches for the creation of digital editions (e.g. graph technology). Furthermore, the partners develop open source, reusable, and largely generic publishing frameworks, i.a. *SADE* (BBAW, SUB)*, ediarum.WEB* (BBAW)*, Wolfenbütteler Digitale Bibliothek*[184] (HAB) and print publications, e.g. *bdnPrint*[185] (SUB) or *ediarum.PDF* (BBAW).

[177] https://xtriples.lod.academy/
[178] https://correspsearch.net/
[179] https://www.bbaw.de/bbaw-digital/telota/forschungsprojekte-und-software/abgeschlossene-projekte/sade
[180] https://dare.uni-koeln.de/
[181] https://lod.academy/site/tools/digicademy/speedy
[182] https://www.ediarum.org/
[183] https://textgrid.de/
[184] http://www.hab.de/de/home/bibliothek/digitale-bibliothek-wdb.html
[185] https://bdn-edition.de/bdnprint.html

Teaching, training and workshops. The partners offer education on digital editing skills in a variety of ways. They actively offer or participate in BA and MA level courses[186] and have been responsible for organising and conducting workshops for more than two decades.

Involvement in working groups and standardisation bodies. The partners participate in the *AG e*Humanities*[187]* of the Union of the German Academies of Sciences and Humanities, the working groups *Research Software Engineering*[188], *Newspapers and Journals*[189], *Digital Publishing*[190], and *Data centres*[191] of the DHd-association, the working group *Digital Data Collections and Text Corpora*[192] of the initiative *Digital Information* of the Alliance of German Science Organisations, the *International TUSTEP User Group*[193], and the *Institute for Documentology and Scholarly Editing*[194]. Further, they are also engaged in a number of standardisation bodies: the *TEI consortium[195]*, the *DTA Base Format*[196], the *IIIF consortium*[197], the *Dublin Core Governing Board*[198], and the *MODS Editorial Committee*[199].

Software sustainability. Partners are often *ex officio* responsible for both the operation of the respective tools they provide and the web portals of digital editions that they maintain and as such have broad experience with sustainability issues. This brings to Text+ a wide range of expertise in conceptual strategies for solving or mitigating these problems.

Repositories and archiving. The partners are experienced in the operation of research data repositories, including long-term archiving (e.g. bitstream preservation), persistent identifiers (e.g. DOIs), and certification processes (e.g. CTS[200], DINI[201], nestor[202]).

5.3.4 Measures

Taking into account the funding cuts, the work programme was adapted accordingly. The available person-months for the implementation reflect not only the cut of 29,8% of the budget, but also the fact that the individual personnel costs increase over the five years of funding, and are higher than the DFG personnel cost rate of 2020, which was the basis for the budget calculation of the proposal. This

[186] AdWMZ at JGU; NRWAW at UniK; BBAW at FUBerlin and HUBerlin; HAB at TUBraunschweig; DACo at TUDA; SUB at UniGOE.
[187] https://www.akademienunion.de/arbeitsgruppen/ehumanities/
[188] https://dh-rse.github.io/
[189] https://dhd-ag-zz.github.io/
[190] https://dig-hum.de/ag-digitales-publizieren
[191] https://dhd-ag-datenzentren.github.io/
[192] https://www.allianzinitiative.de/fields-of-action-projects/digital-data-collections-and-text-corpora/?lang=en
[193] http://www.itug.de/
[194] https://www.i-d-e.de/
[195] https://tei-c.org/
[196] http://www.deutschestextarchiv.de/doku/basisformat/
[197] https://iiif.io/
[198] https://dublincore.org/groups/governing-board/
[199] https://www.loc.gov/standards/mods/editorial-committee.html
[200] https://www.coretrustseal.org/
[201] Deutsche Initiative für Netzwerkinformation e.V., https://dini.de/
[202] Kompetenznetzwerk digitale Langzeitarchivierung, https://www.langzeitarchivierung.de.

means that in terms of funded person-months the reduction for some partners will be significantly higher than 29.8%.

All measures and most of their tasks are essential both for the objectives of the task area Editions as well as the overall Text+ objectives and will be addressed in this funding period, but with limited scope and depth in comparison to the original planning. A lower degree of completeness and integration will be achieved, only a fewer number of outreach activities including e.g. workshops, guidelines and reports will be delivered.

Specifications and requirements will be further developed in the first months of the funding period. This involves in particular own contributions from senior researchers, funded positions will be filled step by step.

In particular, the funding cuts implied the following modifications of the work plan:

- Measure 1.1: the release of the registry will be postponed by six month.
- Measure 1.2: the enhancement of the registry with data from existing catalogues will be limited to selected contents of the envisaged catalogues.
- Measure 1.3: the cycles for specifying “model editions” will be reduced to two.
- Measure 2.1: reports to the SCC will remain annual, for this is of key importance for the interaction with the SCC.
- Measure 2.2: the task area will offer minimum necessary preservation procedures, but will no longer claim to perform full curation of all data sets. The first publication of the documentation of an onboarding policy for this task area will be postponed by 12 months (M24). A revised version will be published at the end of the fourth year (M48).
- Measure 2.3: the development of rescue scenarios for “orphaned editions” will be omitted.
- Measure 2.4: Deliverable E2.4 will be postponed by 6 months
- Measure 3.1: Deliverable E3.1 will be postponed by 6 months
- Measure 3.2: Deliverable E3.2 will be postponed by 6 months
- Measure 3.3: aims at integrating consulting workflows with the consortium’s helpdesk and will no longer include active harmonisation of consulting practices between centres. The report on common consulting workflows will be omitted accordingly.
- Measure 4.1: the offer for coordinated consulting services will be postponed by six months.
- Measure 4.2: training activities and workshops will still be offered on an annual basis, but reduced in duration and scope.
- Measure 4.3: curricular recommendations will be postponed by 12 months.

- Measure 4.4: the collection of existing tutorials will be reduced to the ones provided by the participants in the context of workshops and training activities. The integration into the curated software platform will be limited to two cycles (M24 and M48).
- Measure 5.1: the implementation of a mentoring programme will be omitted.
- Measure 5.2: the release of the curated software platform will be postponed by 6 months. The platform will be reduced to a basic overview of available tools and services.
- Measure 5.3: the enhanced version of the platform ist postponed by 12 months.

Measure 1: Reference Implementation (M1)

To make a substantial, comprehensive, and inclusive set of digital editions accessible, all partners in the Task Area Editions contribute to a curated registry. The registry serves as a knowledge-base to support creators of digital editions by giving structured access to the vast number of editions in the WWW (user story Delivering Support for Digital Edition Projects). The availability of their metadata in the Text+ Discovery Services significantly improves the findability (see M1 Findability in Task Area Infrastructure/Operations) and accessibility (M2 Accessibility in Task Area Infrastructure/Operations) which furthers the recognition of edition projects in the community. The registry is a core instrument for this Task Area. It is based on the work and findings of M3 Standardisation Activities as well as informing the development of the consulting practices carried out in M4 Community Activities. It builds the foundation to identify "model editions" which can serve as best practices for other edition projects in terms of their digital methods. In addition, it supports the decisions for the annual work programme, e.g. by identifying both gaps and synergies in the portfolio (see M2 Portfolio Development). This measure comprises three deliverables:

(1) Release of the registry. Based on a common data model, the registry records information about digital scholarly editions, such as the discipline as well as the language and writing system of the source documents. In addition, it gathers information on editorial guidelines, PIDs, the status of development and maintenance, and the compliance with FAIR and CARE principles and RDM standards. The registry allows for the integration of metadata from both digital and printed editions, so that ideally no editorial heritage is lost (user story A standardized metadata format for the Leibniz Edition).

(2) Enhancement of the registry. Besides existing lists and catalogues (e.g. Franzini203, Sahle204, GEPRIS, AGATE), work on the registry also concerns data on the availability of digitised editions and

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

[204] http://digitale-edition.de/

online versions from discipline-specific services (FIDs)205, as well as other relevant discipline-specific and/or international compilations.206

(3) Iterative specification of "model editions". Based on the guidelines for best practices and quality assurance developed in M3 Standardisation Activities, model editions are identified that represent exemplary editions for specific disciplines, genres, in different languages and scripts, or exemplary applications of a certain editorial method.

Funded Partners (Measures 1–3): AdWMZ, BBAW, DACo, HAB, Leopoldina, NRWAW, STI, SUB, MWS

Deliverable	E1.1	Release of the registry; metadata and API specifications	M30
Deliverable	E1.2	Enhancement with data from existing catalogues	M48
Deliverable	E1.3	Iterative specification of "model editions"	M36, M54
Milestone	E1.4	Substantial and balanced set of digital editions	M60

Measure 2: Portfolio Development (M2)

The aim of this Measure is to establish a clear path to onboarding and integration of resources and expertise into the Task Area Editions. For the initial phase of Text+, the portfolio is distributed among a number of already established and recognised competence centres that offer resources, services, tools, infrastructure, consulting, and training. The portfolio will be continuously extended and complemented based on the criteria for integration or refusal (see Data Selection Criteria in section 5) and the decisions of the SCC, and further resources, services, and expertise from other institutions will be integrated. This Measure develops policies of prioritisation and a scalable approach towards integration in order to accommodate as many resources as possible. It is fundamental for meeting Objective 2 (see section 2.2) for this Task Area. This Measure comprises three deliverables:

(1) Annual monitoring report to the SCC. The registry established in M1 Reference Imple- mentation serves as a monitoring tool. Based on the then-current portfolio and on the registry, the SCC will receive an annual report, which helps the SCC to identify gaps in the portfolio and to evaluate more easily where additional support is needed to achieve compliance with the curation standards (see M3 Standardisation Activities). The registry also helps to identify "orphaned" editions which are no longer maintained and thus are in serious danger of disappearing. Information on endangered material would be the base for developing mitigating measures in a potential second phase of the consortium.

(2) Onboarding policy and scalable approach towards integration. The partners carry out the onboarding procedures and provide comprehensive user guides and documentation for collaborators and data depositors, including well-documented strategies for both the long-term archiving of the

[205] https://www.text-plus.org/en/about-us/further-partners/

[206] e.g. Archivum Medii Aevi Digitale https://www.amad.org/; Mediaevum http://www.mediaevum.de/; Handschriftencensus www.handschriftencensus.de; Porter (2013).

research data and the sustainable maintenance of the respective presentation layers. The partners implement a service portfolio with minimum necessary preservation procedures (i.a. metadata description and registration, separation of data and presentation layer, preservation of underlying data).

(3) Rescue scenarios for precarious and "orphaned" editions. (omitted)

(4) Operational model. The Measure contributes to the development of an operational model (see section 3.5) with regard to the specifics in the Task Area Editions, e.g. the complexity of preservation of individualised presentation layers and the great number of small editorial projects that lack stable institutional ties.

Funded Partners: see Measure 1

Deliverable	E2.1	Annual monitoring report to the SCC	annually
Deliverable	E2.2	Documentation of onboarding policy and procedures for Editions	M24
Deliverable	E2.3	-	
Deliverable	E2.4	Operational model	M54
Milestone	E2.5	Implement onboarding procedures, extended portfolio	M60
Deliverable	E2.6	Sustainability report for Portfolio Development	M60

Measure 3: Standardisation Activities (M3)

Quality standards, interoperability, and sustainability of data and software are crucial for the development and maintenance of digital editions (see the following user stories: (1) Digital Critical Editions and Collections on philosophical authors, (2) Intermedial Editions, (3) Inscriptions in Germany from the Middle Ages to Early Modern Times, (4) Digital Humanities and Medieval Literary Studies, and (5) Literary editions using the example of "Poetry of the German Middle Ages"). To advance the standardisations in these fields and in view of Objective 5 (see section 2.2) this Measure comprises four deliverables:

(1) Quality. Text+ provides guidelines for quality assessment and assurance of digital editions, which will be continuously curated (see user story How to improve historians' skills in digital editing) and concern several aspects: data quality as related to the FAIR principles; quality in terms of data modelling in line with editorial standards; quality of the informational architecture, and quality of the documentation (scientific and technical). The guidelines serve as a reference for the creators of digital editions and will frame the definition of model editions (see M1 Reference Implementation).

(2) Interoperability. Text+ fosters the collaboration between the editing communities and the institutions that manage authority files (e.g. GND) through a workshop and a subsequent report. This deliverable strongly complies with research priorities of communities of interest (see the following user stories: (1) Authority files in digital scholarly editions of correspondence, (2) ZEDAKA – Jewish Welfare and Social Policy, and (3) Distributed editing of 19th century correspondences). Authority files for the explicit and unambiguous identification of persons and other entities are a crucial requirement

for making heterogeneous Humanities data FAIR. Their application allows for linkage of entities within a single edition, as well as for connecting data across editions (see user story Data security, system openness, networking) and other resources (including collections and lexical resources). In addition, the edition projects themselves are keen to actively contribute to the creation and improvement of authority files by providing information obtained through editorial processes. Text+ carries out this deliverable with the DNB and the SUB as coordinating partners and with regard to national and international standardisation initiatives. This will happen in co-operation with the service Metadata, Authority data, Terminologies in M3 Interoperability and Re-usability in the Task Area Infrastructure/Operations.

(3) Sustainability. Text+ elaborates best practices and guidelines for sustainable software development. Since digital scholarly editions at their core are (based on) software, it is of the utmost importance to preserve those software systems alongside with the primary research data. Text+ aims at harmonising the different technical strategies pursued by different institu- tions to mitigate the costs of sustaining long-term availability of research software and digital editions. The strategies include, among other things, centralised hosting of software systems with common requirements and the shared use of system resources, the homogenisation of software stacks, and standard-compliant modelling of system topologies. Text+ will establish cross-cutting Measures with research software engineering activities within the NFDI (e.g. NFDI4Culture and NFDI4RSE).

(4) Consulting workflow. To effectively disseminate outcomes of this Measure and to coordinate the consulting activities among the partners, the centers in TA Editions will integrate their consulting workflows with the consortium's helpdesk.

Funded Partners: see Measure 1

Deliverable	E3.1	Launch of guidelines for quality assurance	M30
Deliverable	E3.2	Workshop and report on interoperability through authority files	M36
Deliverable	E3.3	-	
Deliverable	E3.4	Publication on harmonisation of software stacks	M54
Milestone	E3.5	Standards for quality, sustainability, interoperability, consulting	M60
Deliverable	E3.6	Sustainability report for Portfolio Development	M60

Measure 4: Community Activities (M4)

Text+ conducts a comprehensive set of community activities to provide holistic support for all types of editions and all phases of the edition process. The activities comprise consulting, workshops and training, as well as networking events (user story Delivering Support for Digital Edition Projects), online-tutorials and documentation. They aim at the editing community as a whole (e.g. topics such as legal issues or the usage of authority data), but also include tailor-made offerings (e.g. on non-Latin scripts) for specific target groups (e.g. junior researchers, research software engineers), and different

levels of experience. This Measure will be carried out in collaboration with IO-M4 Community Activities and comprises four deliverables:

(1) Consulting services. Text+ offers coordinated, individual, and specialised case-by-case consulting, with the partners of the network Editions providing the required expertise and experience.

(2) Workshops and training. Text+ offers a wide array of workshops and training (see user stories How to improve historians' skills in digital editing and Digital Humanities at the HAB), both integrated into existing events such as the Edirom or IDE summer schools and carried out as individual offers for specific target groups, including advanced training for staff in crucial positions. Workshops and training activities have a key role in facilitating the dialogue between primarily analogue and digitally driven research. This Measure will foster transdisciplinary cooperation and exchange, including neighbouring fields such as musicology, i.a. Music Encoding Initiative (MEI)-based editions (see Objective 3 Transdisciplinary co-operation).

(3) Curricular recommendations. Text+ transfers the insights gained through the workshops and training into the curricula of educational programmes on scholarly editing. Together with the insights and requirements from the community events, the outcomes are also incorporated into the international discourses (see international network activities listed in 5.3.2 Added Values and 5.3.3 The Network Editions and its Clusters).

(4) Online tutorials. Text+ creates a collection of existing online tutorials on digital editing topics and tools that will be continuously updated, expanded, and promoted. Additionally, new tutorials are created for tools covered by the curated platform developed in M5 Software Services, based on feedback from workshops and training.

Funded Partners: AdWMZ, BBAW, Leopoldina, MWS, NRWAW, DACo, UniPB, HAB, STI.

Deliverable	E4.1	Offering coordinated consulting services	M12
Deliverable	E4.2	Training activities and workshops	annually
Deliverable	E4.3	Curricular recommendations on scholarly editing	M48
Deliverable	E4.4	Collection of existing tutorials, development of new tutorials	M24,M48
Milestone	E4.5	Comprehensive consulting, training and teaching scheme	M60
Deliverable	E4.6	-	

Measure 5: Software Services (M5)

The aim of this Measure is to provide comprehensive software services that support the Humanities community in their engagement with software tools, both through individual guidance and permanently available documentation. This comprises two deliverables:

(1) Mentoring programme. (omitted)

(2) Curated software platform. Information on existing software tools, ranging from small applications to full-fledged editing and publishing environments, is bundled and offered via a curated and commented platform, reflecting the experiences made in the consulting processes. The platform

complements the registry (see M1 Reference Implementation) by providing information, e.g. on application areas, usage contexts (i.e. digital edition projects), supported languages and writing systems, maintainer, available versions, licenses, platform- compatibility, programming languages, and documentation (including tutorials elaborated in M4 Community Activities).

(3) Continuous enhancement. The platform continuously integrates software tools from Editions, the other Text+ domains (e.g. tools for linguistic analysis as demanded, for instance, by user story Supporting Information Retrieval in and for Multilingual Scholarly Editions/Text Resources) as well as from third parties.

Funded Partners (domain-specific work): BBAW, DACo, NRWAW, AdWMz.

Funded Partners (cross-domain work): GWDG

Deliverable	E5.1	-	
Deliverable	E5.2	First release of the curated software platform	M24
Deliverable	E5.3	Enhanced version of the software platform	M48
Milestone	E5.4	Curated tool platform	M60
Deliverable	E5.5	-	

In-kind contributions: AdWMZ, BBAW, DACo, HAB, Leopoldina, MWS, NRWAW, SUB, STI, UniPB.

5.4 Infrastructure/Operations

5.4.1 Motivation for the Task Area Infrastructure/Operations

The Task Area **Infrastructure/Operations** is central in enabling Text+ to support the core usage patterns of an infrastructure addressing language- and text-based research data in the Humanities, such as re-use, development, preservation, or machine processing of research data (see **Grand Challenge II**, section 2.1). Furthermore, the diversity of data, the distributed nature of repositories and infrastructures, and the multitude of service requirements call for coordinated and integrated measures. To achieve this, Infrastructure/Operations pursues the following objectives:

- Providing a FAIR- and CARE-compliant platform of generic services – ranging, e.g., from PID services to generic search and schema management services;
- Connecting the existing and evolving data and service portfolios of the data domains to this platform and among each other using, e.g., standards and authority data;
- Supporting the linkage and integration of other resources and services in the NFDI and beyond;
- Offering easy access to the Text+ service and data portfolio through well-designed user interfaces and APIs;
- Providing means for a reliable and sustainable operation of the infrastructure.

The generic services are based on results of existing infrastructures such as CLARIAH-DE and infrastructure services by the Academies and the DNB (see in particular section 3.1 for information

regarding the background of Text+). Infrastructure/Operations integrates and consolidates the existing and future generic services into a unified platform, puts into place a common technical backbone for a federation of geographically distributed data repositories and clusters, and defines policies and processes for the (unfunded) operation of the generic services through the data centres affiliated with Text+. The fulfilment of these objectives reduces, as a side effect, redundancies within the service portfolio and promotes standardisation and interoperability.

As the different data domains contribute through their individual service portfolios to the overall Text+ service portfolio (see section 4.1), it is of the utmost importance to connect them to the platform of generic services. This is an ongoing process towards the greater goal of a sustainable NFDI. In this context, Infrastructure/Operations provides on-demand service development resources and, in addition, coordinates agile development teams and Measures across the different Task Areas under the umbrella of the cross-cutting Measure M5 Software Services. Furthermore, Infrastructure/ Operations takes care of the service lifecycle management in close coordination with the Text+ Governance.

Easy access to the data and service portfolio is essential for the success of Text+ and the NFDI in general. Infrastructure/Operations takes this into account by providing a web-based portal as the entry-point to the Text+ portfolio. As part of the cross-cutting Measure M4 Community Activities, the development of the portal guarantees a high degree of integration, a unified user interface and user experience, and single sign-on (SSO) to the portal. A helpdesk system, support, and complementary Measures from other Task Areas also contribute to the realisation of the *easy access* paradigm.

The data resources provided by the three data domains and those added during Text+ via any of the four usage patterns (see **Grand Challenge II**, section 2.1) are the core assets of Text+ (and, again, of the NFDI in general). This implies that proper curation wherever possible as well as data and metadata quality assessment and assurance are essential tasks to be carried out by Infrastructure/Operations and by all data domains through the cross-cutting data quality and curation activities (e.g. Measure IO-M3 Interoperability and Re-usability).

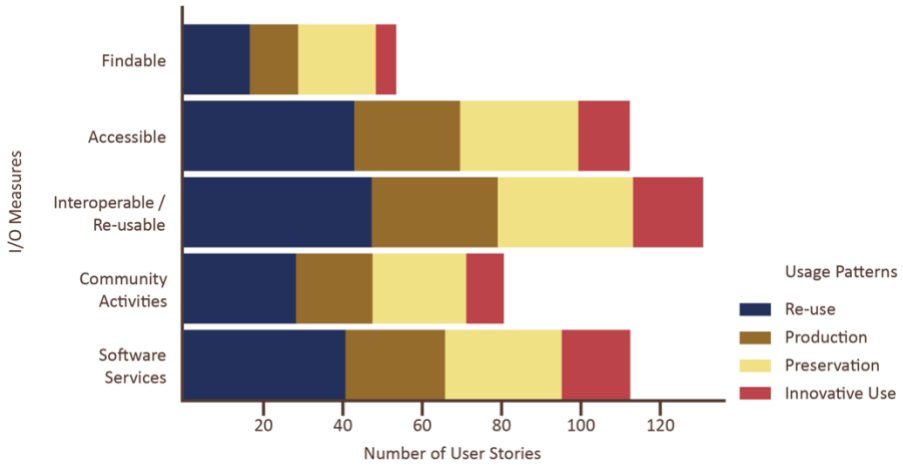

Figure 5.2 Number of user stories assigned to Measures in Infrastructure/Operations, broken down by usage patterns (including multiple assignments)

By design, the Measures in this Task Area support the usage patterns addressed in **Grand Challenge II**. The Measures are also closely related to the user stories. Figure 5.2 illustrates the connection between the user stories, the Measures in the Task Area Infrastructure/Operations, and the usage patterns. Even at this very rough overview level, it becomes apparent that all Infrastructure/Operations Measures address a significant number of use cases and cover the different usage patterns. Moreover, the Measures are designed to support the further development of the forms of use and to react flexibly to newly emerging forms of use.

5.4.2 Added Value of the Task Area Infrastructure/Operations

Various aspects point towards the added value of a centralised Task Area Infrastructure/Operations in Text+. To start with, the Task Area's important role inside Text+ is building the common technological basis for the data domains. In mathematical terms, common services are placed outside the brackets. This reduces redundancies, fosters interoperability and standardisation, and reduces efforts through re-use. For example, the Task Area provides comprehensive search solutions and increases the findability of Text+ resources. A unified user experience, well-defined technical interfaces and quality assurance for data and services are other aims. Furthermore, sustainability and scalability are facilitated with a coordinated technical base infrastructure.

In the NFDI, the Task Area Infrastructure/Operations is a contact point for cross-cutting technical concerns. The Measures in this Task Area assure interoperability and linkage using standards and providing flexible import and export formats. Here, the provision and the use of authority data plays an important role. Finally, the Task Area fosters national and international connectivity by means of appropriate interfaces and through the active participation in relevant networks. These added values from a bird's-eye view will be complemented with more detailed aspects when describing the Measures of the Task Area in section 5.4.4.

5.4.3 Network Infrastructure/Operations

Partners contributing initially to the Task Area are: German National Library (DNB), *Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen* (GWDG), Göttingen State and University Library (SUB), Jülich Supercomputing Centre (FZJ), Leibniz Institute for the German Language (IDS), Saxon Academy of Sciences and Humanities (SAW), Technical University Dresden (TUDD), University of Bamberg (UniBA). The specific competences they bring to the Task Area are described on the Text+ webpage.[207]

5.4.4 Measures

The Measures in this Task Area will support the Clusters in all Text+ objectives (see section 2.2). They will provide services and consultancy and they will foster, among other aspects, consistency and interoperability within Text+ and beyond. The Measures will allow Text+ users to exploit the full potential of the infrastructure. To this end, five Measures will be implemented. Figure 5.3 gives an overview of the Measures and their tasks. The first three Measures in Infrastructure/Operations follow a different design than in the data domain Task Areas and correspond to the FAIR data principles: M1 Findability, M2 Accessibility, and M3 Interoperability and Re-usability. They are complemented, in parallel to the data domain Task Areas, by M4 Community Activities and M5 Software Services.

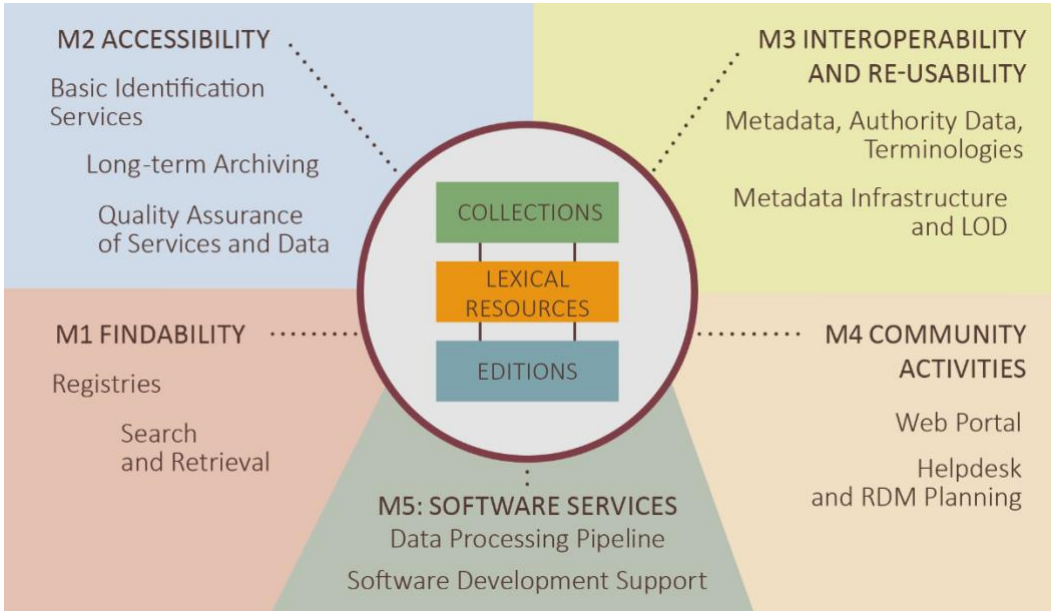

Figure 5.3 Overview of Measures and Tasks in Infrastructure/Operations

For a start, Text+ will rely on the substantial work and experiences that the consortium has gained in, e.g., CLARIAH-DE, the project that merges CLARIN-D and DARIAH-DE, and via the German Academies. On this basis, the Measures will adapt and enhance the services and tools for the requirements in Text+

[207] https://www.text-plus.org/en/research-data/data-and-competence-centres/

and the NFDI. Continuous monitoring and evaluation of the services and tools provided by the Task Area Infrastructure/Operations will be carried out by the OCC. The OCC will discuss and decide on new developments and new proposals for services in the Task Area Infrastructure/Operations based on the evaluation of the existing services, the requirements stemming from the data domains, and technical developments. One guideline in this respect will be the service lifecycle, which will be adapted and further developed for Text+ in Measure M5.

Taking into account the funding cuts, the work programme was adapted accordingly. The available person-months for the implementation reflect not only the cut of 29,8% of the budget, but also the fact that the individual personnel costs increase over the five years of funding and are higher than the DFG personnel cost rate of 2020, which was the basis for the budget calculation of the proposal. This means that in terms of funded person-months the reduction for some partners can account for up to 40%.
All measures and most of their tasks are essential for the Text+ objectives and will be addressed in this funding period, but with limited scope and depth in comparison to the original planning. A lower degree of completeness and integration will be achieved, only a fewer number of outreach activities including e.g. workshops, guidelines and reports will be delivered, and all working cycles are set to 15-month cycles.
Specifications and requirements will be further developed in the first months of the funding period. This involves in particular own contributions from senior researchers, funded positions will be filled step by step. Task plans will be evaluated and potentially revised in 15-month cycles.

In particular, the funding cuts implied the following modifications of the work plan:

- Measure 1, Task (iii) and (iv): implemented interfaces and exchange mechanisms of the federated registry to connect to other initiatives within and beyond the NFDI context will be limited to standardized and established variants
- Measure 1, Task (vi) Design and implementation of a hybrid search and retrieval (SR) architecture will consider partners in the NFDI only
- Measure 1, Task (v) and (xi) Expansion and customization of existing functionalities for project- and institution-specific collections, presentations and search solutions will focus primarily on user interface customizations (i.e. theming capabilities)
- Measure 1, Task (xii) IDS will focus on the FCS integration of KorAP. I.e. it is still planned to integrate Kontext (NoSketchEngine) but only if time permits.

- Measure 1, Task (xiv) and (xv) are merged and focus for improvements in the web-based user interfaces on suitable means to overcome limitations when querying access-restricted resources such as AAI support or context restrictions.
- Measure 2 will in general be restricted to offering the Task (i) to (v) core services, integration of data domain requirements & services, and support. Even there, resources may be limited and decisions will be made (by the OCC) with respect to the requests to be fulfilled. The evolution of the core services based on the requirements of Text+ is not possible anymore. Text+ will therefore bring in its requirements in discussions and potential proposal endeavours related to the topic of "Basisdienste". However, it has to be noted that it is very likely that the best practices processes as proposed initially can not be implemented.
- Measure 2, Task (vi) Support for repository certification is removed. It is envisaged that certification will be supported by Text+ on NFDI e.V. level, and that cross-NFDI support will be established.
- Measure 3, Tasks (i)-(vii) are closely related and will focus on the most important topics requested by the community. The number of workshops and guidelines to be published is being reduced. Text+ will seek for synergies especially with the partner consortia from the humanities and cultural studies.
- Measure 3, Task (viii) must be restricted to selected support of researchers and research projects in using the GND, according to the cut of person-months
- Measure 3, Task (x) designing of an organisational and technical framework, processes, and potentially a platform is removed from the workplan. DNB and SUB will potentially work on a separate funding proposal through in-kind contributions.
- Measure 3, Task (xi): integration of automatic schema matching approaches will be reduced to mature and openly available approaches and strives for a (semi-)automatic mechanism for specialist users to set up mappings between schemata
- Measure 3, Task (xii) and (xiii) the integration with other approaches and normative sources will focus on established, standardized and readily available variants within the NFDI context
- Measure 4 will in general be restricted to offering the core services for supporting the data domains in their community activities.
- Measure 4, Task (i) for connecting services to the webportal, on one hand specifications for the design and technology will be provided for core services that lead to close integration while other services will be loosely integrated and will presumably deviate from the Text+ corporate design in terms of look-and-feel
- Measure 4, Task (ii) implements the basic operation of a helpdesk and general support for tool-based RDM plans. Text+ will support common approaches on NFDI e.V. level and seek for

synergies with other NFDI consortia, especially with the partner consortia from the humanities and cultural studies.

- Measure 5, Task (i) offers on-demand development capacity in the Task Area Infrastructure/Operations in cooperation with all three Data Domains.
- Measure 5, Task (iii) “Text+ Service Infrastructure Best Practices” will be reduced to a basic overview of general infrastructure concepts and best practices. Advanced, detailed documentation of specific services and solutions is not possible anymore.
- Measure 5, Task (iv) “Service Life-cycle Management” will be conducted in collaboration with the Text+ office. Measure 5, Task (iv) will only provide the technical information required for the decision-making processes.

Measure 1: Findability

Dealing with heterogeneous data of different resource types, such as collections, lexical resources, or editions, is the key challenge for findability. Subject-specific requirements demand flexible search services on the one hand and procedures for data integration and interoperability on the other. Starting from Grand Challenge II, two typical usage perspectives with respect to findability will be considered for the Text+ infrastructure: (i) data production and (ii) data re-use.

From the data production perspective, users offer resources via Text+ or add resources to the Text+ infrastructure. To support this requirement, Text+ provides registry services that offer various functionalities to connect resources to the infrastructure, to get an overview of existing resources, and to organise individually relevant sets of resources. This allows researchers to register their resources and to provide them with the required (resource-level) metadata. As an additional service, entry to the registry is also integrated into the Text+ service for creating data management plans (M4 Community Activities). Technically, the registry will be an extended and integrated version of the Virtual Collection Registry (VCR) from CLARIN-D and the Collection Registry from DARIAH-DE.

Considering the data re-use perspective, Text+ offers search services that support users at different levels when they search for and within the resources to find relevant data and tools. On the level of registries, registry entries for all data domains can be searched (on the resource-level metadata; resources being collections, lexical resources, and editions). On the second level (element-level metadata), a flexible Discovery Service working primarily on the element metadata (and, if available and accessible, possibly also on full texts) is offered. On the third level, a more in-depth, content-based search is offered that is specifically designed for annotated linguistic resources. It is planned to extend this search service in the course of Text+ to address, e.g., lexical resources as well.

For the Discovery Service, the administration and indexing of Text+ resources as well as the linkage to connected sources are essential. Due to the overall constellation, the architecture has to take into

account the use of a dedicated index as well as the federation with other search functionalities. With regard to external connectivity, intensive use and support of standard interfaces (e.g. OAI-PMH or SRU/CQL) will be provided. The discovery services will allow handling of schema heterogeneity where necessary, with mappings defined in the Text+ metadata infrastructure (IO-M3 Interoperability and Re-usability). A query defined in one metadata schema can be translated into other schemata if appropriate – e.g. in order to address a broader range of collections.

Besides metadata search, search will also be possible based on the (structured) content of the artefacts as a content-based search (in close co-operation with the Task Areas Collections and Lexical Resources). Based on preliminary work of CLARIN-D and CLARIAH-DE, the Federated Content Search (FCS) provides federated search functionalities on resources such as corpora and treebanks by retrieving data via interfaces from distributed heterogeneous text resources. It allows users to simultaneously retrieve textual data from different providers of resources, while supporting complex queries using linguistic features. It thereby provides an entry point to the broad landscape of corpora engines and query interfaces of the resource providers in the consortium and beyond.

Added value. The Measure ensures findability of all data and metadata, unifies content and presentation of these data, and offers the opportunity of a branded project registry and search tools. Users can define their own compilations of resources and, for example, bundle them under the identity of a project or institution. By using this feature, projects and institutions can form sets of resources that can be used with their own branding (colour scheme, logo, structure of the search results page, etc.) as project resources (branded repository, branded search). In this way, project or institution related resources can easily be maintained using the services of Text+ and at the same time be integrated seamlessly into websites of projects or institutions.

Improvements on findability will allow for a search on different levels of granularity, representation of resource types and their provenance as well as legal aspects for access-restricted data. The Measure thus addresses all Objectives of Text+, and in particular Objective 5. A user-friendly and easy-to-use UI will help to define queries and to retrieve results in a well-designed and comprehensive way.

Target groups/users. The new platform will enable new user groups, research approaches and established data formats in the text-oriented Humanities and Social Sciences to benefit from the Text+ research infrastructure. All researchers of scientific communities targeted by Text+ can use the findability-related services as powerful entry points to all available resources. Scholars and projects that want to administer and provide their collections, lexical resources, or editions are supported by the additional benefit of a branded project registry and search. Resource providers offering access-restricted resources or resource types/formats not yet supported by the FCS, will benefit from the FCS extensions.

Tasks. For all tasks, a search solution via standardised interfaces will be used, which can access distributed resources with the help of metadata harvesting or meta-search concepts. This also enables the connection to other services and infrastructures such as EOSC, Europeana, OpenAIRE, and others. Within the framework of European co-operation, common experiences in the Virtual Language Observatory, the Federated Content Search, and the Generic Search from CLARIN-D and DARIAH-DE will be used.

The following adaptations and extensions will be realised in the Text+ Collection Registry: (i) Coordination with the tool for creating data management plans and provision of appropriate interfaces. (ii) Adaptation to the collection-level metadata schemata in Text+. (iii) Further development to a federated registry in coordination and connection with other NFDI initiatives. (iv) Interfaces and exchange mechanisms to existing and emerging registries beyond the NFDI context. (v) Expansion of the existing elementary functionalities for project- and institution-specific collections and presentations (branded registry, branded search, etc.).

Discovery service functionality will be based on standard retrieval technology (Lucene) and the predecessor solution from CLARIAH. Based on this, the following extensions will be worked on:

(vi) Design and implementation of a hybrid search and retrieval (SR) architecture which considers partners in the NFDI and internationally.

(vii) Integration of additional standard interfaces (e.g. SRU).

(viii) Seamless integration of textual resources of different formats such as editions and lexical resources.

(ix) Improved user-friendliness concerning query formulation (term suggestion, authority data, faceted search, parameter adjustment), and result presentation (expressiveness of results and transparency of the ranking).

(x) Flexible treatment of legal and contractual restrictions.

(xi) Improved customisability for project- and institution-specific search solutions.

For the FCS, the following objectives will be pursued: (xii) Integration of additional local query engines into the FCS requires an adapted strategy based on the complexity of the respective local data model or resource type. Role models for integration are the corpus platform KorAP and the corpus query interface Kontext(NoSketchEngine). (xiii) Evaluation of potential extensions of the FCS specification for improved support of additional resource types. The integration of lexical resources (see M5 Software Services in Task Area Lexical Resources) will serve as a blueprint. (xv) Suitable means to overcome limitations when querying access-restricted resources such as AAI support or context restrictions will be investigated and implemented. (xvi) Evaluation of suitable candidates for integration of standard file formats (such as CoNLL or popular TEI application profiles).

Funded Partners: IDS, SAW, UniBA

Deliverable	IO1.1	Specifications/requirements analysis for tasks i, iii – vii, x, xiii, xvi	M15
Milestone	IO1.2	Extended Working prototypes for tasks ii, iii – vii, xi, xiv	M30
Milestone	IO1.3	First iteration of improved applications/portals based on user feedback for task ii – xii, xiv, xv	M45
Milestone	IO1.4	-	
Milestone	IO1.5	Finalisation of all applications/portals for all tasks	M60
Deliverable	IO1.6	Sustainability report for Findability services	M60

Measure 2: Accessibility

Providing access to Text+ data and services requires technical as well as organisational efforts, which are subsumed under the Measure Accessibility.

Granting access to users, research data, and other research objects is key to a distributed research infrastructure such as the one provided by Text+. The respective Authentication and Authorisation Infrastructure (AAI) for users and Persistent Identifier (PID) services for research data (and other research objects such as software) are the foundations for identification, definition of roles, single sign-on (SSO), rights management, findability, and re-usability. The Text+ AAI is based on the CLARIAH-DE AAI, which already serves the German and European DARIAH and CLARIN communities and will be integrated into the Text+ portal (see M4 Community Activities). The PID services, which are also central to realising Findability, are based on ePIC PIDs as well as on DOIs, which are, in conjunction with the AAI, an integral part of the DARIAH and CLARIN infrastructures. Both services provide globally resolvable identifiers for data and other research objects. While the ePIC PID service is offered by GWDG, the SUB is tasked with a central DOI registration service for the Humanities in Germany. Text+ thereby offers free-of-charge identifier registrations for the Humanities, either by uploading the data sets to the DARIAH repository (see also below), or by providing identifiers for data available through already established data repositories. Both organisations will continue to offer free-of-charge identifier registrations for Humanities data either by allowing uploads of the data sets to the DARIAH repository (see also below), or by offering identifier registration of data available through already established data repositories.

Following the steps of finding (meta)data (see M1 Findability) and authorising access, the actual data objects can be accessed. The fundamental function that is required and implemented by this Measure is the provision of (long-term) access to metadata and the data objects themselves. To achieve this, two different kinds of service are needed: repositories and a long-term archive (LTA). The repositories, which provide online access to research data, are a core service of each thematic cluster. In addition, Text+ offers the DARIAH Repository (see section 4.1) as a “catch-all” repository for data owners who want to benefit from the Text+ offers and community connections and comply with the generic Text+ data selection criteria, but do not fulfil the cluster-specific criteria (see Table 5.1 Overview of Task Areas). Regarding LTA, Text+ will base its service on the koala long term archiving software, which

constitutes the foundation of the CTS- and nestor-certified long-term archive operated by the DNB. It is developed and maintained by the GWDG, which also operates the respective service for the DNB and other organisations. The existing koala service is able to archive large amounts of data in the range of multiple petabytes and the related metadata, verify them, and store them securely.

Complementing this Task Area's range of services, the Accessibility Measure addresses a variety of quality and usability aspects ranging from service availability over trustworthiness and long-term sustainability to monitoring and evaluation of service usage. Service monitoring is a prerequisite for every complex technical infrastructure to ensure quality requirements such as availability, performance, standard compliance, or user acceptance. It is therefore a precondition for providing reliable, performant, and useful services and to evaluate an infrastructure's current state as well as its development over time based on performance indicators. Finally, repository and archive certification address the trustworthiness of data stores. Certification at regular intervals by renowned authorities such as CTS or nestor helps to ensure the reliability, trustworthiness, and persistence of repositories, and demonstrates to users as well as to funders that the infrastructure meets international standards.

Added value. The AAI offers single sign-on (SSO) into the Text+ infrastructure and direct access to all services through the Text+ portal. It enables service providers to easily integrate their services into the Text+ portfolio and to manage their usage. International federations and the integration of the AAI with multiple authentication providers such as eduGAIN allow users to access educational services worldwide. The Text+ AAI is extremely scalable and open to all interested NFDI consortia. The identifier services can generate an unlimited number of PIDs and DOIs and will be tightly integrated into the Text+ infrastructure. These services, too, are open to all interested NFDI consortia.Text+ delivers a distributed and integrated set of certified repositories. Furthermore, it offers a concise long-term archiving policy based on the extensive experience of the DNB and other partners, and the respective Text+ Archive service, which is offered to all users of Text+ (data production) as well as to those who preserve already existing data (data preservation). Due to its generic nature, Text+ Archive is also capable of serving other NFDI initiatives based on their particular long-term archiving policies.

Monitoring and web analytics provide event- and performance-based information on services as well as means for project managers and funders to review the current status of the infrastructure regarding availability, responsiveness, or compliance with other agreed-upon criteria. Furthermore, the gathered data also serves as a foundation for user studies and helps to identify shortcomings in the service portfolio. Certification processes are a driver for adopting and implementing standards. As such, certification requires expert knowledge and guidance. This will ensure the reliability, durability, and trustworthiness of data repositories for both the infrastructure's operators and users.

The Measure supports the implementation of all objectives of Text+ but provides in particular services and support for Objective 5 (see section 2.2).

Target groups/users. The Accessibility services target the communities of interest directly, in particular AAI and repositories. PID services and LTA, although accessible by end users, target mainly data and service providers. Monitoring and certification mainly target repository, storage, and service providers. All services can be used by other NFDI consortia.

Tasks. This Measure will (i) assist service developers with the integration of their services with the Text+ AAI including support, infrastructure customisation, and the ongoing evolution of the AAI according to the needs of the community and the NFDI in general. Moreover, (ii) the PID services will be integrated as the central PID provider of Text+ within the first two years. (iii) Access to the DARIAH Repository will be integrated into the Text+ portal and (iv) the Text+ Archive service will be added to the service portfolio within the first two years of the project. The underlying policy and the service itself will be updated continuously and APIs for all aforementioned services will be developed and published. With respect to (v) monitoring, new service providers, new OAI-PMH endpoints, or new interfaces will be added according to the needs of the providers, and dashboards and reports will be customised.

Funded Partners: DNB, GWDG, SUB

Deliverable	IO2.1	Specification of required adaptation of AAI for Text+, documentation for service developers	M12
Deliverables	IO2.2	Text+ long-term archiving policy	M12
Deliverable	IO2.3	Report on Text+ monitoring requirements with respect to service quality, user satisfaction and integration into Text+	M12
Deliverable	IO2.4	-	
Milestone	IO2.5	Technical monitoring of all public Text+ services available	M24
Milestone	IO2.6	Text+ PID services as central identifier provider in operation	M24
Milestone	IO2.7	Text+ Archive in operation	M24
Milestone	IO2.8	-	
Deliverable	IO.2.9	-	
Milestone	IO2.10	APIs for AAI and PID services publicly available	M30
Deliverable	IO2.11	-	
Milestone	IO2.12	Implemented AAI adaptations in operation	M36
Deliverable	IO2.13	-	
Deliverable	IO2.14	Final report of user satisfaction and service quality	M60
Deliverable	IO2.15	Report on archived resources in Text+ Archive	M60
Deliverable	IO.2.16	Sustainability report for Accessibility services	M60

Measure 3: Interoperability and Re-usability

Implementing the FAIR data principles is at the heart of both the Text+ Task Areas Collections, Lexical Resources, and Editions, and the NFDI as a whole. The use and – where necessary – the (further) development of data standards, metadata formats, authority data, and terminologies are key requirements for interoperability and re-usability. Text+ therefore implements a cross-cutting service Metadata, Authority data, Terminologies which includes the following offers:

First, standardisation and best practices. This includes the promotion of existing standards, the development of guidelines and best practices and, where necessary, supports the community of

interest in developing appropriate standards, data models, and terminologies for their specific needs. Ideally, this leads to standardised application-specific specialisations of already existing standards – here referred to as Application Profiles (AP), including terminology recommendations. An example for this within the TEI framework is DTABf.

Second, a framework for metadata quality management. This includes a metadata assessment tool helping data curators and researchers to form a data improvement strategy as well as consulting and support for adapting the quality analysis to specific needs.

Third, with respect to the use and development of authority data and terminologies, the SUB implements in close co-operation with the DNB a GND Agency for language- and text-based research data (iii), an innovative service which extends the scope of the GND in alignment with the DNB's GND strategy. The GND is aligned with a growing number of national and international vocabularies such as the Standardthesaurus Wirtschaft (STW), ORCID, and Wikidata, and the usability of the GND encompasses all three data domains addressed in Text+ as well as numerous other NFDI initiatives' research data domains. The Measure addresses the integration of Text+-relevant terminologies with the GND and expands its connectivity with other community resources. The GND Agency will represent the requirements of Text+ in the Committee of Library Standards and the GND committee and is open to serve as a liaison for the NFDI to these committees.

In order to best support the tasks mentioned above at a technical level, a sophisticated management of schema and authority file information is required (iv). This involves in-depth knowledge of the individual schemata used as well as mappings between schemata and the linkage to authority files and knowledge graphs. Based on substantial know-how concerning metadata infrastructures and schemata (see section 4.2), Text+ will provide services that allow for expressive modelling and exploitation of individual metadata schemata. The extensive knowledge about schemata and their interrelationships can then be used for deep vertical search solutions but is also essential for data integration and re-use. Furthermore, virtual collections are made possible, through which data collections – e.g. collections of data from different institutions – are created and used for research. Furthermore, the schema management also supports the implementation of application profiles and quality constraints.

The schema management services and tools will also provide means to link to other resources in various ways (v). One important resource in this respect are knowledge graphs (KGs) and collaborative terminology services developed in NFDI initiatives such as NFDI4Culture. A first way of linking is the inclusion of references to KG-nodes in the metadata. This linking will be supported by Text+ search and recommendation techniques. Another important aspect is that the integration of Text+ data in KGs will be supported by export services for respective formats (RDF) which allow to drastically reduce the manual effort for integrating Text+ metadata into a KG. Finally, information coming from KGs can be

integrated in Text+ metadata, e.g. by using view definitions of KG management systems or deploying SPARQL-queries to enrich metadata records using the DME.

Target groups/users. Developers and management of the Text+ infrastructure reliant on the interconnection between Text+ and external partners; users benefitting from additional resources available via Text+ and the advanced quality and interoperability; data producers benefitting from enhanced interoperability and simplified integration procedures.

Added value. This Measure ensures coherence across the Text+ data domains. It delivers contributions to standardisation and best practices in terms of data curation, data quality, and the implementation of the FAIR principles. This includes producing guidelines, supporting data providers in implementing data quality improvements, actual FAIRification of relevant datasets, and the provision of respective services and tools.

The achieved metadata quality is, among other aspects, important for interoperability as well as for the development and automatic configuration of tools and services. Explicit schema mappings allow for the definition of relationships to external schemata and facilitate data exchange and synchronisation. The metadata infrastructure supports the seamless integration of additional resources and resource types. This includes broadening the overview of resources beyond Text+ and the provision of tools to unify, harmonise, and enrich metadata inventories.

An advantage of the Measure's design is that standardisation efforts, consulting, and tool development work in close co-operation to assure a well-aligned service portfolio in consultation with the data domains and Text+ users. The Measure addresses all five Text+ objectives (see section 2.2), especially Objective 3.

Tasks. Creation of a metadata infrastructure for interoperability and re-usability based on standards, integrated authority files, linked open data, schema management services, and tools for data integration and linkage. This includes the following activities:

Organising workshops for scholars and data curators on various levels (from beginners to experts). Possible subjects could be:

(i) a general introduction to metadata, semantic web, interoperability issues, and authority files (i.e. GND);

(ii) extending existing or developing new controlled vocabularies, i.e. ontologies or thesauri, and mapping to other metadata formats and/or external authority data (e.g. GND);

(iii) developing domain specific metadata APs that are compliant with the Singapore Framework;

(iv) using CLARIN's Component Metadata Infrastructure and DARIAH's Data Modeling Environment.

Publishing guidelines and best practice reports on

(v) how to find existing metadata standards, APs and controlled vocabularies and how to re-use them;

(vi) how to conceptually and technically map between metadata standards, APs and controlled vocabularies, how to create a metadata AP for a specific research context;

(vii) how to document and publish metadata APs using the Text+ tools and services.

Establishing an official GND agency for language- and text-based research data which includes:

(viii) support of researchers and research projects in using the GND, in the semi-automatic enrichment of existing data sets with GND links, and in adding new entries or data elements to the GND;

(ix) evaluating new requirements from the community and mediating them with the standardisation bodies;

Based on existing infrastructure components, the following extensions for the metadata infrastructure will be developed:

(xi): integration of automatic schema matching approaches as a proposal mechanism for specialist users to set up mappings between schemata;

(xii) integration and coordination with approaches in other NFDI initiatives and at the European level (e.g. with Gerdi, EUDAT and EOSC);

(xiii) improved integration of lexical resources, specialist sources, and authority data to further improve the validity of schema information and mappings;

(xiv) improved integration of Linked Open Data and authority data with an emphasis on easy-to-use interfaces;

(xv) description of schemata and mappings based on the architecture and services of the Text+ data modelling environment.

Funded Partners: DNB, SAW, SUB, UniBA

Milestone	IO3.1	Two workshops on selected subjects from i – iv	M15, M30, M45, M60
Deliverable	IO3.2	One publication / revised publication on selected subjects from v – vii	M15, M30, M45, M60
Deliverable	IO3.3	Schema descriptions and mappings, progress report (xiv)	M15, M30, M45, M60
Deliverable	IO3.4	Specifications/requirements analysis for tasks viii – xiii	M15
Deliverable	IO3.5	Report on requirements for a GND Agency for language- and text-based research data	M27
Milestone	IO3.6	Stable versions of viii, x and first prototypes of the other tasks	M30

Milestone	IO3.7	Service for semi-automatic enrichment with GND links available	M36
Milestone	IO3.8	GND Agency for language- and text-based research data established	M45
Milestone	IO3.9	Stable versions of ix, xi, xii, xiii and revised versions of viii, x	M45
Milestone	IO3.10	-	
Milestone	IO3.11	Revised and optimised versions of viii – xiii	M60
Deliverable	IO3.12	Sustainability report for Interoperability and Re-usability services	M60

Measure 4: Community Activities (M4)

Text+ offers support and consulting at all stages of the data lifecycle and for all services and tools provided by the consortium. It provides expertise from the three data domains for helpdesk, consulting, and training. Easy access and the possibility to get in personal contact are important prerequisites for a trustworthy community service that is close to the user. Text+ implements the following major offerings: (i) a portal and (ii) a helpdesk including RDM plans, as well as (iii) coordination of community activities.

Offer i) includes the set-up, maintenance, and continuous development of a Text+ portal, which is the central point of entry for users. It integrates access to the helpdesk, comprehensive information on standards, best practices, community activities, as well as the data discovery services and other applications. Offer ii) provides a helpdesk, which allows users to choose where to turn for advice, e.g. for requesting a personal consultation or for receiving support regarding individual tools or offers of Text+. The helpdesk includes support in developing RDM plans for new research projects. While online tools such as RDMO assist in a first step and researchers can also consult with their home institutions, RDM strategies highly depend on the selected research methods of a project. Text+ will provide specialised support for RDM plans.

Finally, offer iii) coordinates all community-related activities of the data domains (M4 Community Activities in Task Areas 5.1-5.3) in order to provide a consistent and coordinated set of training materials and offers. It will analyse helpdesk requests and collect data on both requested and conducted consulting and training activities based on a common procedure for service quality management and improvement purposes. These data can serve as Text+ user input to the portfolio development and, together with further web analytics (see M2 Accessibility), reduce possible thresholds and constantly help to improve the offers.

Target groups/users. Researchers at all stages of their career and from all disciplines addressed in Text+.

Added value. Text+ community services enable researchers to use and benefit from the Text+ services and deliver continuous support, consulting, and training to the community of interest. The Measure thus addresses all Objectives of Text+ (see section 2.2) and has particular impact on Objective 5.

Tasks. The Text+ community services can largely extend existing offers and will be set up through upscaling and integration of the respective services of CLARIN-D and DARIAH-DE, which are currently being merged in CLARIAH-DE. Yet the integration of additional services and the continuous evolvement is a complex challenge, e.g. from a user experience design perspective for the portal and helpdesk, and for the sustained maintenance of the community network.

Funded Partners: GWDG, SUB

Milestone	IO4.1	Initial version of the Text+ website is set up	M1
		Initial version of Text+ portal is set up (i)	M18
Milestone	IO4.2	Initial version of Text+ helpdesk is set up (ii)	M6
Deliverable	IO4.3	Guidelines for specifying RDM plans for text- and language-based research projects	M30
Deliverable	IO4.4	Report on analysis of helpdesk, consulting, and training activities	M18, M36, M54
Deliverable	IO4.5	Sustainability report for Community Activities	M60

Measure 5: Software Services (M5)

Text+ implements the cross-cutting Measure Software Services to realise the consistent development and integration of IT infrastructure, services, APIs, tools, and applications, as well as to offer expertise and processes for requirements engineering, service design and development. The Measure provides the foundation for modern software development based on clearly communicated principles, thus fostering interoperability and sustainability. This project-internal service offer is complemented by a framework for customisable data processing pipelines. Furthermore, Software Services coordinates service development across all Task Areas and provides a Text+ internal development team.

Added value. This Measure delivers “on demand” development effort to assist service providers and application developers with the integration of their software and services through agile teams. Support for the integration of tools and services targeting non-Latin scripts will create further added value. The data processing pipelines will enable the data domains to build innovative research services based on a common, customisable framework. Moreover, this Measure delivers and implements an ITIL-compliant service lifecycle management for Text+ and provides the continuous integration and continuous delivery infrastructure. This Measure supports in particular the data-related Task Areas with the realisation of Objective 2, and Objective 5 (see section 2.2).

Target groups/users. The resources of Software Services are provided mainly to the Text+ data domain-related Task Areas, but also more generally to developers interested in using or integrating Text+ services, including other NFDI consortia.

Tasks. The Task Area Software Services comprises four major Tasks: (i) provision of on-demand development capacity, (ii) provision and customisation of data processing pipelines, (iii) development of guidelines, best practices, and training on how to engage with the Text+ service infrastructure, and

(iv) management of the service lifecycle management process. Task (i) supports data owners, infrastructure experts, service developers, and end users to form dynamic and short-lived teams across Task Areas to implement a particular service, integration, or API. Through Task (ii), the GWDG, JSC, and TUDD deliver an innovative service to improve the way researchers work with data. Instead of setting up their own infrastructure and downloading large amounts of data, researchers stay within the Text+ infrastructure, access the data they need, and get customised processing pipelines. The task will build upon existing, highly scalable, opensource services from experienced academic data and computing centres. This allows for the inclusion of further centres and the integration with other NFDI consortia. Task (iii) defines and communicates the respective best practices and guidelines, propagates sustainability into all architectural levels, and ensures connectivity across the NFDI and internationally. Finally, task (iv) constantly integrates new, innovative services, which are proposed by the Text+ community as well as third parties. The service lifecycle management process, which is based on the DARIAH Service Life Cycle, ensures both sustainability and acceptance through a formal process that a service must pass in order to be successfully integrated into the Text+ research infrastructure.

The necessary foundations for the service development, i.e. processes, guidelines, and best practices, will be developed early on in the project. Over the course of the project, this Measure then takes over core integration and implementation tasks as well as customisation of data processing pipelines based on the Text+ service portfolio management strategy and steered by the respective committees' decisions. This will be done in close collaboration with the Measures M5 Software Services of the data domain-related Task Areas Collections, Lexical Resources and Editions.

Partner: GWDG, JSC, TUDD

Deliverable	IO5.1	Text+ software development guidelines	M12
Deliverable	IO5.2	Text+ service life cycle management guidelines	M6
Milestone	IO5.3	Continuous integration and delivery infrastructure is available	M12
Milestone	IO5.4	Training on Text+ service integration prepared	M18
Deliverable	IO5.5	Training on Text+ service integration prepared	M18
Deliverable	IO5.6	Report on service and portfolio development	annually
Milestone	IO5.7	Text+ code sprint for Humanities data took place	annually from year 2
Deliverable	IO5.8	-	
Milestone	IO5.9	(Updated) Text+ data processing pipeline concept and framework available	M24, M48
Deliverable	IO5.10	(Updated) Text+ data processing pipeline concept and framework available	M24, M48
Deliverable	IO5.11	Sustainability report for Software Services	M60

In-kind contributions: DNB:, GWDG, IDS, JSC, MWS, SAW,, SUB,, TUDD, UniBA

5.5 Administration, Collaboration and Sustainability

The IDS will act as applicant institution for Text+ and will thus take the lead for this Task Area. The IDS will be responsible for the disbursement of project funds and for the day-to-day operation of the

Scientific Office of Text+. The Infrastructure/Operations Office of Text+ and its day-to-day operation will be in the hands of the co-applicant SUB.

The Task Area Administration, Co-operation, and Sustainability will be responsible for (i) the disbursement of funds to partner institutions of Text+; (ii) the disbursement of flexible funds; (iii) Scientific Office and the Infrastructure/Operations Office of Text+; (iv) all coordination tasks for the three SCCs and for the OCC; (v) all coordination tasks concerning cross-cutting topics and (vi) concerning sustainability with the NFDI Directorate. These responsibilities correspond one-to-one to the six Measures defined below.

The Scientific Office and the Infrastructure/Operations Office will cover a wide range of scientific tasks that require advanced knowledge of the scientific, technical, and infrastructural domains of Text+. The spokespersons of the Scientific Office (Erhard Hinrichs) and of the Infrastructure/Operations Office (Regine Stein) will engage in sustained activities at both the national and the international level and will need to rely on continuous support by experienced and highly knowledgeable coordinators. Accordingly, each of the two offices needs to be supported by a full-time post-doc position at a highly advanced experience level.

Coordination of the SCCs and of the OCC will be taken on by the Task Area Administration, rather than by the individual Task Areas described in sections 5.2-5.5. This centralised approach has two distinct advantages for the effective governance of Text+: (i) the Scientific and the Infrastructure/Operations Offices will have a comprehensive overview of the state-of-affairs in all three SCCs and in the OCC at all times and can synchronise their actions; (ii) the support and coordination actions for the three SCCs and for the OCC will be offered at a consistent level of service throughout the funding period. In view of the broad range of tasks (see A-M4 and A-M5 for more details), coordination of the three SCCs requires one full-time and coordination of the OCC one half-time academic position.

The financial administration will require a full-time staff position of a financial officer as well as part-time (0.25 full time equivalent) secretarial support.

5.5.1 Measures

Taking into account the funding cuts, the work programme was adapted accordingly. The available person-months for the implementation reflect not only the cut of 29,8% of the budget, but also the fact that the individual personnel costs increase over the five years of funding and are higher than the DFG personnel cost rate of 2020, which was the basis for the budget calculation of the proposal. This means that in terms of funded person-months the reduction for some partners can account for up to 40%.

All measures and most of their tasks are essential for the Text+ objectives and will be addressed in this funding period, but with limited scope and depth in comparison to the original planning. A lower

degree of completeness will be achieved, and all working cycles which originally were planned for 12-month intervals are now set to 15-month cycles, and the flexible funds are reduced by 32%.

Specifications and requirements will be further developed in the first months of the funding period. This involves in particular own contributions from senior researchers, funded positions will be filled step by step. Task plans will be evaluated and potentially revised in 15-month cycles.

In particular, the funding cuts implied the following modifications of the work plan:

- Measure 2, Deliverable A2.1: Due to the extended cycle of 15 months for the flexible funds, the number of deliverables has been reduced. Furthermore, the amount of flexible funds is reduced by 32%.
- Measure 3
    - Deliverable A3.1: the meetings of the scientific board no longer take place on a quarterly basis but rather on a four-month cycle. Accordingly, the reporting is adjusted to this longer cycle.
    - Deliverable A3.3 and A3.4 the reports will not be independent documents but integrated into annual reports to the funders and organizational entities
- Measure 4, Deliverable A4.1: the reports will no longer be quarterly, but rather on a four-month cycle.
- Measure 5, Deliverable A5.1 the report will not be an independent document but integrated into annual reports to the funders and organizational entities
- Measure 6, Deliverable A6.2 is deleted, as bylaws do not match the requirements and regulations of the consortium agreement.

Measure 1: Financial Administration of the Consortium

Financial administration involves the following subtasks to be shared by a financial officer and by secretarial support: disbursement of project funds to a large consortium of co-applicant and participant institutions; continuous monitoring of the total yearly budget and of the yearly expenditures of individual co-applicant and participant institutions; coordination of financial transactions with the funding agency, and financial reporting to the funding agency; reimbursement of travel funds to members of the following committees: SCC, OCC, EB, AB.

Target groups/users. The Text+ consortium.

Added value. This Measure will be in charge of the full financial administration on behalf of the entire Text+ consortium.

This measure is not reduced.

Funded Partners. IDS

Deliverable	A1.1	Monitoring and disbursement of project funds to partners; reimbursement of travel costs to SCC, OCC, EB, AB; financial reporting to the Funding Agency	annually

Measure 2: Administration of Flexible Funds

The Text+ budget includes a pool of flexible funds for extending the Text+ portfolio of data and services. The allocation of flexible funds to additional participant institutions will be determined by the SCCs and the Executive Board. Financial administration involves the following subtasks to be shared by a financial officer and by secretarial support: disbursement of project funds to additional participant institutions; continuous monitoring of the total yearly budget and of the yearly expenditures of flexible funds; coordination of financial transactions with the funding agency, and financial reporting to the funding agency. Due to the reduced funding the yearly cycles will be implemented in four 15-month cycles

Target groups/users. The Text+ consortium.

Added value. This Measure will be in charge of the full financial administration on behalf of the participant institutions that will receive flexible funds.

Funded Partners: IDS (for redistribution)

Deliverable	A2.1	Disbursement of flexible funds and administrative integration of additional participants	M15, M30, M45

Measure 3: Text+ Office: Scientific Office and Infrastructure/Operations Office

The Text+ office is tasked with the preparation of the plenary meetings of Text+, with serving as a point of contact for communities of interest, contributing to trainings and community services and their research projects, and coordinating and maintaining the liaison activities with national and international partner infrastructure initiatives, as well as with technical and administrative coordination between the OCC and the SCCs. In addition, the Scientific Office is solely responsible for the support of the scientific speaker in all day-to-day operations of Text+: co-authoring and coordinating all scientific and administrative reports to the NFDI Directorate, the funding agency, and to the boards of Text+, and coordinating all administrative issues with the NFDI Directorate. The Infrastructure/Operations Office is solely responsible for the support of the operations speaker in all day-to-day operations of Text+: co-authoring and coordinating all Infrastructure/Operations reports, providing a point of contact with the operators of technical infrastructures, and supervising the helpdesk.

Target groups/users. The Text+ consortium.

Added value. This Measure will be responsible for coordinating all tasks of the Scientific Office and the Infrastructure/Operations Office.

Funded Partners: IDS, SUB

Deliverable	A3.1	Preparation of and reporting on meetings (normally every four months) of the Scientific Board	annually
Deliverable	A3.2	Preparation and hosting of the annual Text+ plenary meeting	annually
Deliverable	A3.3	Input for the annual report on liaison activities with national and international partner infrastructure initiatives	annually
Deliverable	A3.4	Input for the annual report on training and dissemination activities	annually

Measure 4: SCC and OCC Coordination

This Measure assists all three SCCs and the OCC, including the respective chairs, with all coordination activities for the SCCs and the OCC. This includes: soliciting and compiling reports from all Measures in the three data domains for the purposes of monitoring progress; preparing agendas for SCC and the OCC meetings; assembling and distributing documents in preparation of SCC and the OCC meetings; and taking minutes and documenting the work of each SCC and for the OCC. Each of the SCCs and the OCC meets four times a year, resulting in a total of 16 meetings per year that need to be coordinated and supported. In addition, organisational support for the chairpersons is provided with respect to all coordination tasks of their SCCs.

Target groups/users. The Text+ consortium.

Added value. This Measure will coordinate all activities of the SCCs and the OCC.

Funded Partners: IDS; SUB

Deliverable	A4.1	Preparation of and reporting on SSC and OCC meetings; administrative support of the SCC and OCC chairs	annually

Measure 5: Co-operation within the NFDI: Cross-cutting topics

Description: This Measure assists Text+ experts from the three data domains and the Infrastructure/ Operations domain in their contributions to cross-cutting topics coordinated within the NFDI as a whole. This assistance consists of soliciting and compiling written input from Text+ experts for specific cross-cutting topics; dissemination of reports of working groups for cross-cutting topics to the Text+ consortium; coordination of NFDI working group meetings; and assisting Text+ experts who chair working groups on cross-cutting topics in their administrative tasks.

Target groups/users. The Text+ consortium, other consortia and the NFDI Directorate.

Added value. This Measure will coordinate all activities of Text+ on cross-cutting topics within the NFDI.

Funded Partners: IDS, SUB

Deliverable	A5.1	Input for the annual report on Text+ contributions to NFDI cross-cutting topics; dissemination of NFDI documents on cross-cutting	annually

	topics		

Measure 6: Co-operation within the NFDI: Sustainability

This Measure will coordinate the necessary steps for ensuring the sustainability of Text+. It involves co-operation of the NFDI directorate with the institutional heads of Text+ (Co-)Applicant Institutions and Participants. The Text+ office will assist in this process by managing the following steps: drafting and signing a consortium agreement for Text+; drafting and signing by-laws for all committees in the governance of Text+; and developing an operating agreement in consultation with all Text+ partner institutions.

Target groups/users. The Text+ consortium and the NFDI Directorate.

Added value. This Measure will harmonise Text+'s process of sustainability planning with the overall sustainability development within the NFDI Directorate.

Funded Partners. IDS; SUB

Deliverable	A6.1	Preparation and signing of the consortium agreement	M 3
Deliverable	A6.2	-	
Deliverable	A6.3	Preparation of the Text+ operating model	M 60

5.6 Risk Management

Text+ will generate and regularly update a risk management register based on the initial risk assessment below.

Objective (see section 2.2)	**Risk**	**Mitigation (measures to prevent risk from occurring)**	**Likeli-hood**	**Impact**	**Contingency Plan**
Support methodological diversity by high-quality Research Data.	Community uptake is limited to a few disciplines and standardised methodologies, which will impede development.	SSC membership is distributed over a broad range of disciplines and organisations, with varying methodological approaches and needs for innovation.	low	high	Intensify collaboration among communities with high level of interest as showcases for widening collaboration.
Comply with research priorities of communities of interest.	The research data offers for inclusion in the Text+ portfolio do not meet FAIR and CARE principles.	Community and consulting services enable data providers to improve the quality of their resources prior to submission.	low	medium	Discuss with the communities and with the NFDI directorate a common strategy for FAIR-ifying and CARE-ifying research data.
	The research data offers for inclusion in the Text+ portfolio exceed the capacities of Text+.	- The SCCs select those data and services that have the highest relevance for the communities of interest. - Text+ experts provide the SCCs with an estimate of the amount of effort involved for resource integration into the portfolio.	high	medium/ low	Develop cost sharing models and encourage data providers to budget these costs in their grant applications so as to generate additional resources for data inclusion.

	Insufficient community take-up of the data-handling standards, procedures and guidelines of Text+.	Offering continuous community and consulting services in close coordination with SCCs and OCC.	high	medium	Seek consultation with the NFDI directorate and with the DFG about formulating a general policy for the take up of the FAIR principles across scientific disciplines and communities.
Foster transdisciplinary cooperation.	Cross-cutting topics do not receive enough support within Text+.	Text+ coordination and management continuously highlight the importance of cross-cutting topics.	low	medium	Seek partnerships with other (NFDI) consortia or software initiatives; shift resources toward community activities.
Advance innovative research.	SSC favour established research methods and overlook novel research directions.	SSC are tasked with tracking emerging types of research data and research methodologies.	medium	medium	Together with the NFDI directorate, Text+ will develop and offer incentives for active participation in innovative research methods and data as well as in cross-cutting topics.
Objective (see section 2.2)	**Risk**	**Mitigation (measures to prevent risk from occurring)**	**Likeli-hood**	**Impact**	**Contingency Plan**
Improve research data on a massive scale.	The resources to be integrated do not meet state-of-art quality standards.	Community and consulting services enable data providers to improve the quality of their resources prior to submission.	high	medium/ low	Shift more resources to community services and consulting services. Seek an even closer coordination with the NFDI directorate to publish guidelines for research data quality.
	Existing resources become technologically outdated or inaccessible or can no longer be maintained.	Include only components that meet sustainability criteria. Avoid proprietary software or data formats. Maintain continuous technology watch.	medium	high	Seek partnerships with other (NFDI) consortia or software initiatives; shift resources toward software development.

Appendix

A. Bibliography and List of References

Sources which were written or developed by members of the consortium are highlighted in bold. All web pages were last accessed on 28.09.2020.

Alliance of German Science Organisations (2010): *Principles for the Handling of Research Data*. Available at: https://www.mpg.de/230783/Principles_Research_Data_2010.pdf.

Arnold, A.; Fisseni, B.; Kamocki, P.; Schonefeld, O.; Kupietz, M.; Schmidt, Th. (2020): Addressing Cha(lle)nges in Long-Term Archiving of Large Corpora. In: **P. Bański et al.** (Eds.), *Proceedings of the LREC 2020 8th Workshop on Challenges in the Management of Large Corpora (CMLC-8)*. 1-9. European Language Resources Association (ELRA), Paris. Available at: https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/CMLC-8book.pdf.

Bański, P.; Kupietz, M.; Witt, A., Ćavar; D., Heiden, S.; Aristar, A.; H. Aristar-Dry (2012) (Eds.): *Proceedings of the LREC 2012 Workshop on Challenges in the Management of Large Corpora (CMLC)*, 22 May 2012, Istanbul, Turkey. European Language Resources Association (ELRA), Paris. European Language Resources Association (ELRA), Paris. Available at: http://www.lrec-conf.org/proceedings/lrec2012/workshops/05.CMLC-Proceedings.pdf.

Bański, P.; Bingel, J.; Diewald, N.; Frick, E.; Hanl, M.; Kupietz, M.; Pezik, P.; Schnober, C.; Witt, A. (2013): KorAP: The New Corpus Analysis Platform at IDS Mannheim. In: Vetulani, Z.; Uszkoreit, H. (Eds.): *Proceedings of the 6th Language and Technology Conference (LTC'13)*. 586-587. Available at: https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3261/file/Banski_KorAP_2013.pdf.

Bierwirth, M.; Glöckner, F. O.; Grimm, C.; Schimmler, S.; Boehm, F.; Busse, C.; Degkwitz, A.; Koepler, O.; Neuroth, H. (2020): *Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung*. DOI: 10.5281/zenodo.3895209.

Bollmann, M. (2019): A Large-Scale Comparison of Historical Text Normalization Systems. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1* (Long and Short Papers). 3885-3898. Association for Computational Linguistics, Stroudsburg, PA. Available at: http://www.aclweb.org/anthology/N19-1389.

Brünger-Weilandt, S.; Bruhn, K.; 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*. DOI: 10.5281/zenodo.4045000.

Brüning, G.; Henzel, K.; Pravida, D. (2013): Multiple Encoding in Genetic Editions: The Case of "Faust". In: *Journal of the Text Encoding Initiative,* (4). DOI: 10.4000/jtei.697.

Cavoukian A. (2009): *Privacy by Design: The 7 Foundational Principles*. Available at: https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf.

Damon, C. (2016): Beyond Variants: Some Digital Desiderata for the Critical Apparatus of Ancient Greek and Latin Texts. In: M. J. Driscoll; E. Pierazzo (Eds.), *Digital Scholarly Editing: Theories and Practices*. 201-218. Open Book Publishers, Cambridge. Available at: https://books.openedition.org/obp/3421.

Deutsche Forschungsgemeinschaft (2015): *Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora. Handreichung*. Available at: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf.

English translation: *Guidelines for Building Language Corpora Under German Law*. translated by **E. Ketzan; J. Wildgans; J. Weitzmann** (2017). Available at: https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_review_board_linguistics_corpora.pdf.

Deutsche Forschungsgemeinschaft (2016): *DFG Practical Guidelines on Digitisation*. Available at: https://www.dfg.de/formulare/12_151/12_151_en.pdf.

Diesner, J. (2019): *Understanding Social Structure and Behavior through Responsible Mixed-Methods Research: Bias Detection, Theory Validation, and Data Governance*. Keynote at the DHd Conference 2019, Frankfurt. Available at: https://dhd2019.org/programm/fr/keynote-jana-diesner/.

Diewald, N.; Margaretha, E. (2016): Krill: KorAP Search and Analysis Engine. In: *Corpus Linguistic Software Tools (= Journal for Language Technology and Computational Linguistics (JLCL) 1/2016)*. 73-90. Available at: https://jlcl.org/content/2-allissues/5-Heft1-2016/jlcl-2016-1-4DiewaldMargaretha.pdf.

Dima, C.; Hinrichs, E. (2015): Automatic Noun Compound Interpretation Using Deep Neural Networks and Word Embeddings. In: M. Purver; M. Sadrzadeh; M. Stone (Eds.): *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*. 173-183. London. Available at: https://www.aclweb.org/anthology/W15-0122/.

Eckart, T.; Gradl, T. (2017): Working towards a Metadata Federation of CLARIN and DARIAH-DE. In: *Proceedings of the CLARIN Annual Conference 2017. n.p. CLARIN ERIC, Utrecht*. DOI: 10.5281/zenodo.1173604.

Firth, J. R. (1957): *Papers in Linguistics 1934-1951*. London, Oxford University Press.

Geyken, A.; Haaf, S.; Jurish, B.; Schulz, M.; Steinmann, J.; Thomas, C.; Wiegand, C. (2011): Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In: S. Schomburg et al. (Eds.), *Proceedings of Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland.* Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz). 157-161. Available at: https://hbz.opus.hbz-nrw.de/opus45-hbz/frontdoor/deliver/index/docId/206/file/PDFA_Tagung_Digitale_Wissenschaft_hbz_2011_7.pdf.

Glöckner, F.O.; Diepenbroek, M.; Felden, J.; Overmann, J.; Bonn, A.; Gemeinholzer, B.; Güntsch, A.; König-Ries, B.; Seeger, B.; Pollex-Krüger, A.; Fluck, J.; Pigeot, I.; Kirsten, T.; Mühlhaus, T.; Wolf, Ch.; Heinrich, U.; Steinbeck, Ch.; Koepler, O.; Stegle, O.; Weimann, J.; Schörner-Sadenius, T.; Gutt, Ch.; Stahl, F.; Wagemann, K.; Schrade, T.; Schmitt, R.; Eberl, Ch.; Gauterin, F.; Schultz, M.; Bernard, L. (2019): *Berlin Declaration on NFDI Cross-Cutting Topics*. DOI: 10.5281/zenodo.3457213.

Gradl, T.; Henrich, A. (2014): A Novel Approach for a Reusable Federation of Research Data within the Arts and Humanities. In: *Digital Humanities 2014 – Book of Abstracts*. 382-384. EPFL – UNIL, Lausanne (Switzerland). Available at: http://dharchive.org/paper/DH2014/Paper-779.xml.

Gradl, T.; Henrich, A. (2016): Data Integration for the Arts and Humanities: A Language Theoretical Concept. In: *20th International Conference on Theory and Practice of Digital Libraries*, TPDL 2016, Hannover. DOI: 10.1007/978-3-319-43997-6_22.

Gray, R.D.; Drummond, A.J.; Greenhill, S.J. (2009): Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. In: *Science*, *323*. 479-483.

Gülden, S. A.; Krause, C.; Verhoeven, U. (2020): Digital Palaeography of Hieratic. In: V. Davies; D. Laboury (Eds.), *The Oxford Handbook of Egyptian Epigraphy and Paleography*. DOI: 10.1093/oxfordhb/9780190604653.013.42.

Hamp, B.; Feldweg, H. (1997): GermaNet: A Lexical-Semantic Net for German. In: *Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. 9-15. Madrid. Available at: http://www.sfs.uni-tuebingen.de/GermaNet/.

Hankinson, A.; Jefferies, N.; Metz, R.; Morley, J.; Warner, S.; Woods, A. (2020): *Oxford Common File Layout Specification. Recommendation. Version 1.0*. Available at: https://ocfl.io/1.0/spec/.

Henrich, V.; Hinrichs, E. (2010): GernEDiT: The GermaNet Editing Tool. In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*. European

Language Resources Association (ELRA), Paris. 2228-2235. Available at: https://www.aclweb.org/anthology/L10-1180/.

Hildenbrandt, V.; **Moulin, C.** (2012): Das Trierer Wörterbuchnetz: Vom Einzelwörterbuch zum lexikographischen Informationssystem. In: *Korrespondenzblatt des Vereins für niederdeutsche Sprachforschung, 119*. 73-81.

HRK (Hochschulrektorenkonferenz) (2016): Wie Hochschulleitungen die Entwicklung des Forschungsdatenmanagements steuern können. Orientierungspfade, Handlungsoptionen, Szenarien. Empfehlung der 19. Mitgliederversammlung der HRK am 10. November 2015 in Kiel. In: *Beiträge zur Hochschulpolitik, Vol. 1/2016*. Available at: https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-10-Publikationsdatenbank/Beitr-2016-01_Forschungsdatenmanagement.pdf.

Horn, F. (2020): *Der Ereignis- und Objektcharakter von Briefen im 19. Jahrhundert: Briefformate und Reflexionen zum Briefempfang digital auswerten*. TUPrints, Darmstadt. DOI: 10.25534/tuprints-00009137.

Horstmann, W.; Nurberger, A.; Shearer, K.; Wolski, M. (2017): *Addressing the Gaps: Recommendations for Supporting the Long Tail of Research Data*. DOI: 10.15497/RDA00023.

Hotson, H.; Wallnig, T. (Eds.) (2019): *Reassembling the Republic of Letters in the Digital Age. Standards, Systems, Scholarship*. Göttingen University Press, Göttingen. DOI: 10.17875/gup2019-1146.

Hulle, D. V. (2016): Modelling a Digital Scholarly Edition for Genetic Criticism: A Rapprochement. In: *Variants, 12-13*. 34-56. Available at: https://journals.openedition.org/variants/293.

Jurish, B. (2011): *Finite-state Canonicalization Techniques for Historical German*. Ph.D. Thesis, Universität Potsdam.

Kamocki, P.; Ketzan, E.; Wildgans, J.; Witt, A. (2018): CLARIN Legal Information Plattformen und Legal Helpdesk. In: G. Vogeler (Ed.), *DHd 2018, Kritik der digitalen Vernunft, Konferenzabstracts*. 365-366. DOI: 10.18716/KUPS.8085.

Kamocki, P.; Stauch, M. (2020): "Cover this data that I cannot see": Privacy by Design in Machine Translation. In: J. Porsiel (Ed.), *Maschinelle Übersetzung für Übersetzungsprofis*. 42-57. Available at: https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/9926.

Kamocki, P.; Witt, A. (2020): Privacy by Design and Language Resources. In: N. Calzolari et al. (Eds.): *Proceedings of the Twelfth Conference on International Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA), Paris. 3423-3427. Available at: https://www.aclweb.org/anthology/2020.lrec-1.419/.

Kett, J. (2017): *Initiative für Normdaten und Vernetzung: GND-Entwicklungsprogramm 2017-2021*. Available at:

https://wiki.dnb.de/download/attachments/132749726/GND_Entwicklungsprogramm17-21_2017-06.pdf?version=1&modificationDate=1516963688000&api=v2.

Király, P. (2017): Towards an Extensible Measurement of Metadata Quality. In: *Conference Proceedings of the Second International Conference on Digital Access to Textual Cultural Heritage*. 111-115. ACM, Göttingen. DOI: 10.1145/3078081.3078109.

Klein, W.; Geyken, A. (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: U. Heid et al. (Eds.), *Lexicographica*. 79-93. De Gruyter, Berlin/New York.

Kline, M.-J.; Holbrook Perdue, S. (2008): *A Guide to Documentary Editing*, Charlottesville. Available at: https://gde.upress.virginia.edu/.

Kukutai, T.; Taylor, J. (Eds.) (2016): *Indigenous Data Sovereignty: Toward an Agenda*. Australian National University Press. DOI: http://dx.doi.org/10.22459/CAEPR38.11.2016

Kupietz, M. (2015): Constructing a Corpus. In: P. Durkin (Ed.): *The Oxford Handbook of Lexicography.* Oxford University Press, Oxford. 62-75. DOI: 10.1093/oxfordhb/9780199691630.013.5.

Kupietz, M.; Belica, C.; Keibel, H.; Witt, A. (2010): The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In: N. Calzolari et al. (Eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*. 1848-1854. European Language Resources Association (ELRA), Paris. Available at: https://www.aclweb.org/anthology/L10-1285/.

Kupietz, M.; Keibel, H. (2009): The Mannheim German Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research. In: M. Minegishi & Y. Kawaguchi (Eds.), *Working Papers in Corpus-based Linguistics and Language Education, No. 3.* 53-59. Tokyo University of Foreign Studies (TUFS), Tokyo. Available at: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf.

Kupietz, M.; Diewald, N.; Hanl, M.; Margaretha, E. (2017): Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. In: M. Konopka; A. Wöllstein (Eds.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung (Jahrbuch des Instituts für Deutsche Sprache 2016)*. 319-329. De Gruyter, Berlin/Boston. Available at: https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/5965.

Kupietz, M.; Diewald, N.; Fankhauser, P. (2018a): How to Get the Computation Near the Data: Improving Data Accessibility to, and Reusability of Analysis Functions in Corpus Query Platforms. In Bański, P. et al. (Eds.), *Proceedings of the LREC 2018 6th Workshop Challenges in the Management of Large Corpora (CMLC-6)*. 20-25. European Language Resources Association (ELRA), Paris. Available at: http://lrec-conf.org/workshops/lrec2018/W17/pdf/book_of_proceedings.pdf.

Kupietz, M.; Lüngen, H.; Kamocki, P.; Witt, A. (2018b): The German Reference Corpus DeReKo: New Developments – New Opportunities. In: N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 4353-4360. European Language Resources Association (ELRA), Paris. Available at: http://www.lrec-conf.org/proceedings/lrec2018/summaries/737.html.

Kupietz, M.; Schmidt, Th. (Eds.) (2018): *Korpuslinguistik.* (= Germanistische Sprachwissenschaft um 2020, 5). De Gruyter, Berlin/Boston. DOI 10.1515/9783110538649.

Lobin, H.; Schneider, R.; Witt, A. (Eds.) (2018): *Digitale Infrastrukturen für die germanistische Forschung.* (= Germanistische Sprachwissenschaft um 2020, 6). De Gruyter, Berlin/Boston. Available at: https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/7678.

Lusetti, M.; Ruzsics, T.; Göhring, A.; Tanja Samardžić, T.; Elisabeth Stark (2018): Encoder-Decoder Methods for Text Normalization. In: *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects* (*VarDial 2018)*. 18-28. Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/W18-3902/.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. (2013): Distributed Representations of Words and Phrases and their Compositionality. In: C.J.C. Burges; L. Bottou; M. Welling; Z. Ghahramani; K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26 (NIPS 2013).* 3111-3119. Lake Tahoe, NV. Available at https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

Mondaca, F.; Rau, F. (2020): Transforming the Cologne Digital Sanskrit Dictionaries into Ontolex-Lemon. In: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*. 11-14. Available at: https://www.aclweb.org/anthology/2020.ldl-1.2/.

Müller-Spitzer, C. (2010): OWID: A Dictionary Net for Corpus-based Lexicography of Contemporary German. In: A. Dykstra; T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress*. 445-452. Fryske Akademy, Leeuwarden.

Nerbonne, J. (2009): Data-Driven Dialectology. In: *Language and Linguistics Compass* 3(1). 175-198. DOI: 10.1111/j.1749-818X.2008.00114.x.

Pierazzo, E. (2014): Digital Documentary Editions and the Others. In: *Scholarly Editing: The Annual of the Association for Documentary Editing, 35*. 1-23. Available at: http://scholarlyediting.org/2014/essays/essay.pierazzo.html.

Porter, D. (2013): Medievalists and the Scholarly Digital Edition. In: *Scholarly Editing: The Annual of the Association for Documentary Editing, 34*. 1-26. Available at: http://scholarlyediting.org/2013/essays/essay.porter.html.

Sahle, P. (2016): What is a scholary digital edition (SDE)? In: M. Driscoll; E. Pierazzo (Eds.), *Digital Scholary Editing. Theory, Practice and Future Perspectives*. 19-39. Open Book Publishers, Cambridge. Available at: https://books.openedition.org/obp/3397.

Schimmler, S. (2020): *National Research Data Infrastructure for Data Science and AI*. Available at: https://www.dfg.de/download/pdf/foerderung/programme/nfdi/nfdi_konferenz_2020/nfdi4datascience_abstract.pdf.

Schöch, Ch.; Döhl, F.; Rettinger, A.; Gius, E.; Trilcke, P.; **Leinen, P**.; Jannidis, F.; Hinzmann, M.; Röpke, J. (2020): Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: *Zeitschrift für digitale Geisteswissenschaften (in print)*.

Schmidt, Th. (2012): EXMARaLDA and the FOLK Tools: Two Toolsets for Transcribing and Annotating Spoken Language. In: N. Calzolari et al. (Eds.), *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Paris. 236-240.
Available at: https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/185.

Schmidt, Th.; Gasch, J.; Kaiser, J. (2018): DGD: Die Datenbank für Gesprochenes Deutsch. In: **L. M. Eichinger; A. Plewnia** (Eds.), *Neues vom heutigen Deutsch. (Jahrbuch des Instituts für Deutsche Sprache 2018)*. De Gruyter, Berlin/Boston. 351-354.

Stehouwer, H.; Ďurčo, M.; Auer, E.; Broeder, D. (2012): Federated Search: Towards a Common Search Infrastructure. In: N. Calzolari et al. (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, İstanbul (Turkey), 3255-3259. European Language Resources Association (ELRA), Paris. Available at: https://www.aclweb.org/anthology/L12-1291/.

Tittel, S.; Chiarcos, C. (2018): Historical Lexicography of Old French and Linked Open Data: Transforming the Resources of the *Dictionnaire Étymologique de l'Ancien Français* with OntoLex-Lemon. In: *Proceedings of the LREC-2018 GLOBALEX Workshop (GLOBALEX-2018)*. European Language Resources Association (ELRA), Paris. Available at: http://lrec-conf.org/workshops/lrec2018/W33/index.html.

Tonne, D.; Krewet, M.; Hegel, Ph.; Götzelmann, G.; Söring, S. (2019): Aristoteles auf Reisen: Handschriftenforschung in der digitalen Infrastruktur des SFB 980 "Episteme in Bewegung". In: M. Huber; S. Krämer; C. Pias (Eds.), *Forschungsinfrastrukturen in den digitalen Geisteswissenschaften: Wie verändern digitale Infrastrukturen die Praxis der Geisteswissenschaften?* CompaRe, Frankfurt a. M., 77-87. Available at: http://d-nb.info/1201549485/34.

Trilcke, P. (2013): Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. In: P. Ajouri; K. Mellmann; C. Rauen (Eds.), *Empirie in der Literaturwissenschaft*. 201-247. mentis, Paderborn.

van Uytvanck, D.; Zinn, C.; Broeder, D.; Wittenburg, P.; Gardelleni, M. (2010): Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In: N. Calzolari et al. (Eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*. 900-903. European Language Resources Association (ELRA), Paris. Available at: https://www.aclweb.org/anthology/L10-1187/

van Zundert, J. (2012): If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities. In: *Historical Social Research / Historische Sozialforschung, 37,* 165-186. Available at: https://www.jstor.org/stable/41636603.

Wallot, S.; Hollis, G.; van Rooij, M. (2013): Connected Text Reading and Differences in Text Reading Fluency in Adult Readers. In: *Plos One, 8(8), e71914.* DOI: 10.1371/journal.pone.0071914.

Weitin, Th.; Werber, N. (Eds.) (2017): Scalable Reading. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 47,1.

Wittenburg, P.; Mosel, U.; Dwyer, A. (2002): Methods of Language Documentation in the DOBES project. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. European Language Resources Association (ELRA), Paris. Available at: https://www.aclweb.org/anthology/L02-1221/.

B. Curricula Vitae and Lists of Publications

PD Dr. Alexander Geyken
b. *07.06.1963*

Education

2017	Habilitation, Linguistics, University of Potsdam. Thesis title: *Die Zukunft allgemeinsprachlicher Referenzwörterbücher*
1998	Ph.D., Computational Linguistics, Centre for Information and Language Processing, University of Munich. Thesis title: *Aufbau einer französisch-deutschen kontrastiven Lexikongrammatik am Beispiel der Kommunikationsverben*

Professional Experience

since 2019	Research Coordinator of the *Zentrum für digitale Lexikographie der deutschen Sprache* (ZDL) at the BBAW (Berlin-Brandenburg Academy of Sciences and the Humanities)
since 2014	Project Director of the CLARIN-D centre at the BBAW
since 2007	Research Coordinator of the long-term project *Digitales Wörterbuch der deutschen Sprache* (DWDS)
since 1999	Researcher at the BBAW
1997 – 1998	Researcher at the Institute of Romance Languages and Literatures, Department of Linguistics, University of Düsseldorf
1995 – 1996	Researcher at the Leibniz Institute for the German Language, Mannheim
1992 – 1994	Researcher at the Institute for Educational Psychology, University of Munich
1990 – 1992	Researcher at the IBM-Paris Centre Scientifique, France

Committees, Boards and Professional Appointments (Selection)

since 2018	Member of the working group *Digital Data Collections and Text Corpora*, Alliance of Science Organizations in Germany (https://www.allianzinitiative.de/handlungsfelder/digitale-datensammlungen-und-textkorpora/)
since 2018	Member of the Advisory Board, FID Germanistik (http://www.germanistik-im-netz.de/gin-beirat.html)
since 2015	Member of the TELOTA steering committee at the BBAW
since 2015	Member of the DHd-Verband (Digital Humanities in the German-speaking World)

2013	Co-organiser of the DFG round table *Empfehlungen für datentechnische Standards und Tools bei der Erhebung von Schriftkorpora*

Selected Projects and Grants

since 2016	Co-Applicant for ZHistLex, funded by the BMBF
since 2015	Co-Applicant and Scientific Coordinator for OCR-D, funded by the German Research Foundation (DFG)
since 2011	Co-Applicant and Scientific Coordinator for CLARIN-D, funded by the BMBF
2007 – 2024	Scientific Coordinator for the *Digitales Wörterbuch* (DWDS), funded as a long-term project
2007 – 2016	Co-Lead for the *Deutsches Textarchiv* (DTA), funded by the German Research Foundation (DFG)
2007 – 2014	Co-Applicant for dlexdb; project awards by the German Research Foundation (DFG)

Selected Publications

a) Peer Reviewed Publications:

Burckhardt, D., A. Geyken, A. Saupe & T. Werneke (2019). Distant Reading in der Zeitgeschichte: Möglichkeiten und Grenzen einer computergestützten Historischen Semantik am Beispiel der DDR-Presse. *Zeithistorische Forschungen*. Vol. 1/2019.

Geyken, A. & S. Haaf (2018). Integration heterogener historischer Textkorpora in das Deutsche Textarchiv. Strategien der Anlagerung und Perspektiven der Nachnutzung. J. Gessinger, A. Redder & U. Schmitz (eds.). *Korpuslinguistik*, (Osnabrücker Beiträge zur Sprachtheorie 92). Duisburg, pp. 175-192.

Geyken, A., F. Wiegand & K.-M. Würzner (2017). On-the-fly Generation of Dictionary Articles for the DWDS Website. I. Kosem, C. Tiberius, M. Jakubícek, J. Kallas, S. Krek & V. Baisa (eds.). *Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 Conference*, Leiden, Netherlands. Brno: Lexical Computing CZ s.r.o., pp. 560–570.

Geyken, A. (2014). Methoden bei der Wörterbuchplanung in Zeiten der Internetlexikographie. U. Heid, S. Schierholz, W. Schweickard, H. E. Wiegand, R. H. Gouws & W. Wolski (eds.). *Lexicographica*. Berlin/New York, pp. 77-112.

Geyken, A. (2013). Large-Scale Documentary Dictionaries on the Internet. R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.). *An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, pp. 1053-1069.

Heister, J., K.-M. Würzner, J. Bubenzer, E. Pohl, T. Hanneforth, A. Geyken & R. Kliegl (2011). DlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*. Vol. 62.1, pp. 10-20.

Geyken, A., J. Didakowski & A. Siebert (2009). Generation of word profiles for large German corpora. Y. Kawaguchi et al. (eds.). *Corpus Analysis and Variation in Linguistics*, (Tokyo University of Foreign Studies, Studies in Linguistics 1), pp. 141-157.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. C. Fellbaum (ed.). *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London, pp. 23-41.

b) Other Publications:

DWDS: A comprehensive lexical information system of dictionaries, corpora and corpus statistics. URL: www.dwds.de / https://www.dwds.de/imprint

dlexdb: A lexical database for linguistic and psycholinguistic research. Compiled and edited by the University of Potsdam and the BBAW. URL: http://www.dlexdb.de/ / http://www.dlexdb.de/imprint/

Prof. Dr. Erhard Hinrichs
b. *17.08.1954*

Education

1985	Ph.D., Linguistics, The Ohio State University, Columbus, Ohio, USA
1985	MA., Linguistics, The Ohio State University, Columbus, Ohio, USA
1981	Staatsexamen, English and Protestant Theology, University of Tübingen

Professional Experience

since 2019	Senior Researcher in the Department of Digital Linguistics/Research Infrastructures, Leibniz Institute for the German Language, Mannheim.
since 1991	Full Professor and Chair of General and Computational Linguistics, Department of Linguistics, University of Tübingen.
2004 – 2010	Associate Dean of Research of the Faculty of Modern Languages and Literatures, University of Tübingen.
2004	Visiting Professor, Department of Linguistics, Stockholm University, Sweden.
1999	Visiting Professor, Department of Linguistics, University of Illinois, USA.
1997 – 2016	Distinguished Consulting Professor for Computational Linguistics, The Ohio State University, Department of Linguistics, USA.
1993 – 1994	Senior Fellow in the Visiting Professor Program of NTT Laboratories, NTT Information Science Laboratory, Yokosuka, Japan.
1989 – 1990	Visiting Professor, Department of Computer Science, Saarland University, Germany.
1987 – 1990	Assistant Professor of Linguistics, University of Illinois, USA.
1987 – 1990	Research Fellow in Cognitive Science, Beckman Institute for Advanced Science and Technology, University of Illinois, USA.
1985 – 1987	Research Scientist in Artificial Intelligence, Bolt Beranek and Newman Laboratories, Cambridge, Massachusetts, USA.

Scholarships, Awards and Administrative Functions

since 2019	Member of the Advisory Board of the KULTSAM Project funded by the Leibniz Gemeinschaft.

since 2019	Member of the International Advisory Board of the CLARIN-DARIAH Bulgarian Infrastructure Project (ClaDa-BG).
since 2017	Member of the Scientific Committee of the Centro interdisciplinare di richerche per la Computerizzazioni dei segni dell'espressione at the University Cattolica Milano.
since 2016	Honorary Life-long Member of the Linguistic Society of America (LSA).
2012 – 2015	Member of the International Advisory Board of the CLARIN-NL Project.
since 2008	National Coordinator of CLARIN-D and Member of the National Coordinators' Forum of CLARIN-ERIC.
2008 – 2016	Member of the Executive Board of the European Research Infrastructure Consortium *Common Language Resources and Technology Infrastructure* (CLARIN-ERIC).
since 2005	Honorary Life-long Member of the Foundation for Logic, Language, and Information (FOLLI).
1999 – 2000	Speaker of the Collaborative Research Centre *Theoretical Foundations of Computational Linguistics* (SFB 340), University of Stuttgart/University of Tübingen.
1997 – 1998	President of the Foundation for Logic, Language, and Information.
1995 – 1997	President of the European Chapter of the Association for Computational Linguistics.
1992 – 1995	Speaker of the DFG-Graduiertenkolleg *Integriertes Linguistik-Studium*.

Selected Research Grants

2010 – 2022	Principal Investigator of Research Grant *Corpus-based Semantic Composition Models for Phrases* and of Research Grant *INF: Heterogenous Primary Research Data of the SFB 833.* Project awards by the German Research Foundation (DFG) in the Collaborative Research Centre *The Construction of Meaning* (Sonderforschungsbereich 833).
2019 – 2021	Principal Investigator of Research Grant *CLARIAH-DE*. Project award by the German Ministry of Education and Research (BMBF).
2019 – 2021	Principal Investigator of Research Grant *Social Sciences and Humanities Open Cloud* (SSHOC). Project award by the European Commission under the Horizon2020 Programme.
2018 – 2020	Principal Investigator of Research Grant *Modelling of Lexical-semantic Relations for Collocations*. Project award by the German Research Foundation (DFG).

2015 – 2017	Principal Investigator of Research Grant *CLARIN PLUS – Common Language Resources and Technology Infrastructure*. Project award by the European Commission under the Horizon2020 Programme.
2011 – 2020	Project Coordinator and Head of the German Section of the CLARIN European Research Infrastructure Federation: *CLARIN-D: A Web and Centres-based Research Infrastructure for the Social Sciences and Humanities*. Project award by the German Ministry of Education and Research (BMBF).
2008 – 2012	Principal Investigator of Research Grant *CLARIN – Common Language Resources and Technology Infrastructure*. Award by the European Commission.
2008 – 2010	Principal Investigator of Research Grant *Language Technology for Lifelong Learning* (LTfLL). Award by the European Commission.
2006 – 2010	Project Coordinator and Principal Investigator of Collaborative Research Grant *BulDialects – Measuring linguistic unity and diversity in Europe*. In co-operation with the Alfa-Informatica at the Rijksuniversiteit, Groningen, The Netherlands and the Bulgarian Academy of Science, Sofia, Bulgaria. Project award by the *VolkswagenStiftung*.
2005 – 2007	Principal Investigator of Research Grant *Language Technology for eLearning* (LT4eL). Award by the European Commission.
2001 – 2003	Project Coordinator and Principal Investigator of Collaborative Research Grant *Medien-intensive Lehre in der Computerlinguistik-Ausbildung* (MiLCA). Project award by the German Ministry of Education and Research (BMBF).
1999 – 2008	Principal Investigator of Research Grant *Repräsentation und Erschließung linguistischer Daten* and of Research Grant *Linguistische Datenstrukturen: sprachübergreifende Annotationen und Datenklassen*. Project awards by the German Research Foundation (DFG) in the Collaborative Research Centre *Linguistic Data Structures.* (Sonderforschungsbereich 441: *Linguistische Datenstrukturen*).
1992 – 2000	Principal Investigator of Research Grant *Development of an HPSG Syntaxfragment for German*, of Research Grant *Constraints on Grammar for Efficient Generation*, and of Research Grant *From Constraints to Rules: Efficient Compilation of HPSG Grammars*. Project awards by the German Research Foundation (DFG) in the Collaborative Research Centre *Linguistic Foundations of Computational Linguistics* (Sonderforschungsbereich 340: *Linguistische Grundlagen für die Computerlinguistik*).
1992 – 2001	Scientific Director (1992-1995) and Principal Investigator (1992-2001) of the Doctoral and Postdoctoral Fellowship Programme (Graduiertenkolleg) *Integriertes Linguistik-Studium* awarded to the University of Tübingen by the German Research Foundation (DFG).

Selected Publications

Dima, C., D. de Kok, N. Witte & E. Hinrichs (2019). No Word is an Island – a Transformation Weighting Model for Semantic Composition. *Transactions of the Association for Computational Linguistics*. Vol. 7, pp. 437-451.

Hinrichs, E. (2018). Digitale Forschungsinfrastrukturen für die Sprachwissenschaft. H. Lobin, R. Schneider, & A. Witt (eds.). *Digitale Infrastrukturen für die germanistische Forschung*. Germanistische Sprachwissenschaft um 2020, Band 6, Institut für Deutsche Sprache: De Gruyter Verlag, pp. 33-52.

Hinrichs, E., N. Ide, J. Pustejovsky, J. Hajic, M. Hinrichs, M. Fazleh Elahi, K. Suderman, M. Verhagen, K. Rim, P. Stranak & J. Misutka (2018). Bridging the LAPPS Grid and CLARIN. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), Miyazaki, Japan, pp. 1294-1302.

Hinrichs, E. & T. Trippel (2017). CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften. *Bibliothek: Forschung und Praxis*. Vol. 41.1, De Gruyter Verlag, Berlin, pp. 45-54.

De Kok, D. & E. Hinrichs (2016). Transition-based Dependency Parsing with Topological Fields. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, Berlin, Germany, pp. 1-7. ACL 2016 Outstanding Paper.

Hinrichs, E. (2016). Substitute Infinitives and Oberfeld Placement of Auxiliaries in German Subordinate Clauses: A Synchronic and Diachronic Corpus Study Using the CLARIN Research Infrastructure. *Lingua*. Vol. 178, pp. 46-70.

Sorokin, D., C. Dima & E. Hinrichs (2015). Classifying Semantic Relations in German Nominal Compounds using a Hybrid Annotation Scheme. *Journal of Cognitive Science*. Vol. 16.3, pp. 261-286.

Dima, C. & E. Hinrichs (2015). Automatic Noun Compound Interpretation using Deep Neural Networks and Word Embeddings. *Proceedings of the 11th International Conference on Computational Semantics* (IWCS 2015), London, UK, pp. 173-183.

Hinrichs, E. & S. Krauwer (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC14). Reykjavik, Iceland, pp. 1525-1531.

Hinrichs, E., V. Henrich & R. Barkey (2013). Using Part-Whole Relations for Automatic Deduction of Compound-internal Relations in GermaNet. *Language Resources and Evaluation. Vol. 47.3,* pp. 839-858.

Dr. Peter Leinen
b. *27.04.1959*

Education

1986 – 1990	Ph.D., Mathematics, University of Dortmund
1979 – 1986	University Degree in Mathematics and Computer Science, RWTH Aachen

Professional Experience

since 2016	Head of Information Infrastructure (Research data strategy, establishing DH projects), German National Library
2011 – 2016	Head of the Computer Center (Building up a community-based research infrastructure within Baden-Württemberg), University of Mannheim
2004 – 2010	Head of the Computer Center (Building up a specific research infrastructure), University of Trier
1990 – 2004	Research in Scientific Computing (Development of methods and parallel software in the field of fluid mechanics, project manager within the special research programme *Methods and Algorithms for Simulating Physical Processes on Supercomputers*), University of Tübingen
1986 – 1990	Research in Scientific Computing (Development of a new paradigm for software development for parallel computers, Development of an Adaptive Finite Element Code), University of Dortmund

Committees

since 2018	Nestor: Speaker of the management board
since 2018	DARIAH-DE/CLARIN-D: Member of the common Technical Advisory Board
since 2017	LIBER: Member of the DH working group
since 2017	IFLA: Member of the Standing Committee Information Technology
since 2016	Competence Network *Deutsche Digitale Bibliothek* (DDB): Member of the General Meeting
since 2016	Committee for Library Standards
since 2016	German Initiative for Network Information (DINI): Executive Board
since 2016	World Wide Web Consortium (W3C): Advisory Committee
2011 – 2018	DARIAH-DE: Member of the Scientific Board

Selected Publications

a) Articles:

Schöch, C., F. Döhl, A. Rettinger, E. Gius, P. Trilcke, P. Leinen, F. Jannidis, M. Hinzmann & J. Röpke (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. Zeitschrift für digitale Geisteswissenschaften (submitted).

Leinen, P. (2020). Die Rolle der Deutschen Nationalbibliothek beim Aufbau einer Nationalen Forschungsdateninfrastruktur. *Dialog mit Bibliotheken 2020/1*. pp. 15-16. https://d-nb.info/1206109068/34

Jannidis, F., L. Konle & P. Leinen (2019). Thematic complexity. DH2019, Utrecht, 2019, https://dev.clariah.nl/files/dh2019/boa/0504.html

Jannidis, F., L. Konle & P. Leinen (2019). Makroanalytische Untersuchung von Heftromanen. *Book of Abstracts of DHd 2019*, Frankfurt am Main 2019, pp. 167-172.

Klingler, M., P. Leinen & H. Yserentant (2005). The finite mass method on domains with boundary. *SIAM J. Sci. Comput*. Vol. 26.5, pp. 1744-1759.

Leinen, P. (2000). Realization of the Finite Mass Method. P. M. A. Sloot et al. (eds). *Computational Science – ICCS 2002*. pp. 470-479.

Gauger, C., P. Leinen & H. Yserentant (2000). The finite mass method. *SIAM J. Numer. Anal.* Vol. 37.6, pp. 1768-1799.

b) Electronic Publications:

Deutsche Initiative für Netzwerkinformationen e.V. (2018), ed. DINI-AG/ZKI-Kommission *E-Framework*: Handreichung zur Entwicklung und Umsetzung von Serviceportfolios zur nachhaltigen Unterstützung der Digitalisierung in Forschung, Lehre, Studium und Verwaltung, http://dx.doi.org/10.18452/19177.

ALWR-BW (2015). Rahmenkonzept der Hochschulen des Landes Baden-Württemberg für datenintensive Dienste – bw DATA (2015-2019).

Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (2014), ed. E-Science: Wissenschaft unter neuen Rahmenbedingungen, Fachkonzept zur Weiterentwicklung der wissenschaftlichen Infrastruktur in Baden-Württemberg https://mwk.baden-wuerttemberg.de/fileadmin/redaktion/m-mwk/intern/dateien/pdf/Forschung/066_PM_Anlage_E-Science_Web.pdf.

ALWR-BW (2012). Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das Hochleistungsrechnen, https://mwk.baden-wuerttemberg.de/fileadmin/redaktion/m-mwk/intern/dateien/pdf/Forschung/Umsetzungskonzept_bwHPC.pdf.

Prof. Dr. Dr. h.c. Andreas Speer

b. *19.06.1957*

Education

1994	Habilitation, University of Cologne, Philosophy
1986	Ph.D., University of Bonn, Philosophy
1976 – 1982	Studies in Philosophy, Theology (Cath.), Classical Philology, Art History 1981 Diploma Theology (Cath.), 1982 MA Philosophy, 1981/2 Staatsexamen in Latin, Theology (Cath.) and Philosophy

Professional Experience

since 2018	Speaker of the *Albertus Magnus Center for Early Career Researchers*
since 2016	Speaker of the *AG eHumanities* of the Union of German Academies
since 2012	Speaker of the *Cologne Center for eHumanities (CCeH)*
since 2012	Director of the *a.r.t.e.s. Graduate School for the Humanities Cologne* (financed by the “Excellence Initiative”)
2008 – 2012	Director of the Graduate School *A.R.T.E.S.* (*NRW-Forschungsschule*)
Autumn 2006	James Collins Visiting Professor, Saint Louis University
2005 – 2012	Speaker of the *Center for Medieval Studies*, University of Cologne
since Spring 2004	Full Professor for Philosophy and Director of the Thomas-Institut, University of Cologne
2000 – 2004	Full Professor for Philosophy, Institute for Philosophy, University of Würzburg
1999	Visiting Professor at the *Hoger Instituut voor Wijsbegeerte* and Research-Fellow at the *DeWulf Mansion-Center*, University of Leuven
1998	Extraordinary Professor for Philosophy, University of Cologne
Spring 1996	Visiting Professor at the Medieval Institute, University of Notre Dame
1995 – 2000	Heisenberg-Scholarship (DFG)
1988 – 1995	Research Assistant, Thomas-Institut, University of Cologne

Honours and Community Service

since 2020	Member of the Senate of the DFG
2016 – 2020	Member and Speaker of the DFG Review Board *Philosophy*
2016	Speaker of the *AG eHumanities* of the Union of the German Academies
2016	Co-Director of the *Averroes Edition* (Academy project)

2015	Elected member of the *European Academy of Sciences* (EURASC)
2013	Elected full member of the *North Rhine-Westphalian Academy of Sciences, Humanities and the Arts*
2011 – 2020	Research and Graduate Dean of the Faculty for Arts and the Humanities, University of Cologne
since 2007	Director of the *Averroes Latinus*, nominated by the *Union Académique Internationale* (UAI)
2005	Dr. honoris causa of the St. Kliment Ochridski-University, Sofia
2004 – 2013	President of the *Gesellschaft für Philosophie des Mittelalters und der Renaissance* (GPMR)
2002	Elected member of the *Akademie gemeinnütziger Wissenschaften zu Erfurt*
1997 – 2007	Board member of the *Société Internationale pour l'Étude de la Philosophie Médiévale* (S.I.E.P.M.)

Editorial Boards

Miscellanea Mediaevalia (General Editor)

Studien und Texte zur Geistesgeschichte des Mittelalters (General Editor)

Recherches de Théologie et Philosophie médiévale (Editorial Board)

RTPM – Bibliotheca (Editorial Board)

Archiv für mittelalterliche Philosophie und Kultur (Editorial Board)

Others

Director of numerous grant funded research projects (ERC-MSCA, *Deutsche Forschungsgemeinschaft*, *Fritz Thyssen Stiftung*, *Alfried Krupp zu Bohlen und Halbach-Stiftung*, *Gerda Henkel Stiftung*, Nordrhein-Westfälische Akademie der Wissenschaften und der Künste, MIUR, etc.)

Organiser and Co-organiser of numerous international conferences, colloquia and lecture series at the Universities of Cologne, Würzburg, Erfurt, Sofia, Lecce, Notre Dame (IN), and at the Herzog August Bibliothek Wolfenbüttel (HAB)

Organiser of the *Kölner Mediaevistentagungen*, *Philosophie kontrovers*, *Albertus Magnus-Professorship*

Referee for the *Deutsche Forschungsgemeinschaft* (DFG), *Fritz-Thyssen Stiftung*, *VolkswagenStiftung*, *Schweizer Nationalfond*, *Alexander von Humboldt-Stiftung*, CINECA / MIUR, etc.

Selected Publications

a) Peer-Reviewed Publications:

Speer, A. (2018). "*qui prius philosophati sunt de veritate …*" Mittelalterhistoriographie im Wandel. A. Speer & M. Mauriège (eds.). *Irrtum – Error – Erreur* (Miscellanea Mediaevalia 40). Berlin/Boston, pp. 783-809.

Speer, A. (2018). Determined Freedom. Thomas Aquinas on Free Choice. A. Beccarisi & F. Retucci (eds.). *Moral Agency and its Constraints: Fate, Determinism and Free Will in the Middle Ages* = Medioevo 42 (2017). Padova, pp. 163-186.

Speer, A. (2017). Blind Spots of Digital Editions: The Case of Huge Text Corpora in Philosophy, Theology and the History of Sciences. *Advances in Digital Scholarly Editing. Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp.* Leiden, pp. 191200.

Speer, A. (2014). Naturgesetz und Dekalog bei Thomas von Aquin. A. Speer & G. Guldentops (eds.). *Das Gesetz – The Law – La Loi* (Miscellanea Mediaevalia 38). Berlin/Boston, pp. 350-370.

Speer, A. (2014). The Durandus Project at the Thomas-Institut: The *Status Quaestionis*. A. Speer, F. Retucci, T. Jeschke & G. Guldentops (eds.). *Durandus and His Sentences Commentary: Historical, Philosophical and Theological Issues* (RTPM – Bibliotheca, 9). Leuven, Paris & Walpole, MA, pp. 71-96.

Speer, A. (2011). The Power of Wisdom: Four Case Studies of a Late Thirteenth Century Debate. J. Canning, E. King & M. Staub (eds.). *Knowledge, Discipline and Power in the Middle Ages*, (Studien und Texte zur Geistesgeschichte des Mittelalters 106). Leiden/Boston, pp. 175-199.

b) Monographs and Edited Volumes:

Speer, A. & L. Reuke (2020), eds. *Die Bibliothek – The Library – La Bibliothèque* (Miscellanea Mediaevalia 41), Berlin/Boston.

Speer, A. & M. Mauriège (2018), eds. *Irrtum – Error – Erreur* (Miscellanea Mediaevalia 40), Berlin/Boston.

Speer, A. (2014), ed. *Zwischen Kunsthandwerk und Kunst: Die ‚Schedula diversarum artium'* (Miscellanea Mediaevalia 37), Berlin/Boston.

Speer, A. (2010), ed. *Fragile Konvergenz. 3 Essays zu Fragen metaphysischen Denkens* (éditions questions Sonderband 7), Köln.

Regine Stein
b. *31.07.1969*

Education

1996	Diploma in Mathematics, University of Marburg

Professional Experience

since 2018	Deputy Head of the Research and Development Department and Head of the Research Infrastructure Program, Göttingen State and University Library, University of Göttingen
since 2018	Co-Head of the DARIAH-DE Coordination Office, DARIAH-DE
2008 – 2018	Head of Information Technology (Deutsches Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg), University of Marburg
1999 – 2008	Researcher, Zuse Institute Berlin
1997 – 1998	Researcher, Institute for Ecological Economy Research, Berlin

Membership and Services

since 2018	Member of the Europeana Data Quality Committee
since 2018	Chair of the CIDOC working group *Data Harvesting and Interchange / LIDO*
since 2017	Speaker of the working group *Data Exchange* (Deutscher Museumsbund)
since 2001	Member of the CIDOC – ICOM International Committee for Documentation

Fundings

2019 – 2022	Kontinuierliches Qualitätsmanagement von dynamischen Forschungsdaten zu Objekten der materiellen Kultur unter Nutzung des LIDO-Standards – KONDA (Federal Ministry of Education and Research)
2019 – 2021	CLARIAH-DE (Federal Ministry of Education and Research)
2019 – 2020	Europeana Archaeology (European Commission)
2019	IIIF 2019 Conference – Göttingen (IIIF International Image Interoperability Framework)

2017 – 2020	KultSam – Kulturelle Sammlungen als digitaler Wissensspeicher für Forschung, Lehre und öffentliche Vermittlung (Federal Ministry of Education and Research)
2016 – 2019	DARIAH-DE – Überführung der digitalen Forschungsinfrastrukturen für die e-Humanities in die Operational Phase (Betriebsphase) (Federal Ministry of Education and Research)
2013 – 2015	Athena Plus – Access to Cultural Heritage Networks for Europeana (European Commission)
2012 – 2014	Partage Plus – Digitising and Enabling Art Nouveau for Europeana (European Commission)
2012 – 2017	Infrastrukturen für ein internationales Netzwerk kunsthistorischer Fotosammlungen auf Basis des LIDO-Standards (German Research Foundation)
2011 – 2013	Linked Heritage – Coordination of Standards and Technologies for the Enrichment of Europeana (European Commission)

Selected Works and Publications

Knaus, G., R. Stein & A. Kailus (2019). *LIDO-Handbuch für die Erfassung und Publikation von Metadaten zu kulturellen Objekten. Band 1*: *Graphik*. Heidelberg: arthistoricum.net. https://doi.org/10.11588/arthistoricum.382.544

Stein, R. & O. Balandi (2019). Using LIDO for Evolving Object Documentation into CIDOC CRM. *Heritage* 2.1, pp. 1023-1031.

Blümm, M., F. Cremer, P. Gietz, L. Klaffki, C. Kudella, B. Mache, R. Stein & C. Thiel (2019). Betrieb des DARIAH-DE Coordination Office. *DARIAH-DE Working Papers Nr. 38*. Göttingen: DARIAH-DE. URN: urn:nbn:de:gbv:7-dariah-2019-10-1.

Kailus, A. & R. Stein (2018). Besser vernetzt: Über den Mehrwert von Standards und Normdaten zur Bilderschließung. *Computing Art Reader. Einführung in die Digitale Kunstgeschichte.* P. Kuroczyński, P. Bell & L. Dieckmann (eds.). Heidelberg: arthistoricum.net.

Simou, N. & R. Stein (2015). *Linking of Metadata to External Data Sources. AthenaPlus Access to cultural heritage networks for Europeana Publication.* Deliverable D4.6. http://www.athenaplus.eu/getFile.php?id=595

Stein, R. & W. Köhler (2013). Review on Linked Open Data Sources. *AthenaPlus Access to cultural heritage networks for Europeana*. Deliverable D4.2. http://www.athenaplus.eu/getFile.php?id=190

Simou, N., E. Tsalapati, N. Drosopoulos & R. Stein (2012). Evolving LIDO based aggregations into Linked Data. CIDOC Conference *Enriching Cultural Heritage*, Helsinki. http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/ConferencePapers/2012/simou.pdf

McKenna, G. & R. Stein (2011). Best practice report on cultural heritage linked data and metadata standards. *Linked Heritage Coordination of standards and technologies for the enrichment of Europeana*. Deliverable D2.1. http://www.linkedheritage.eu/getFile.php?id=229

Coburn, E., R. Light, G. McKenna, R. Stein & A. Vitzthum (2010). LIDO – Lightweight Information Describing Objects Version 1.0. ICOM-CIDOC Data Harvesting and Interchange Working Group.
http://www.lido-schema.org
Specification: http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf
XML Schema: http://www.lido-schema.org/schema/v1.0/lido-v1.0.xsd

Stein, R., J. Gottschewski, R. Heuchert, A. Ermert, M. Hagedorn-Saupe, H.-J. Hansen, C. Saro, R. Scheffel & G. Schulte-Dornberg (2005). Das CIDOC Conceptual Reference Model – Eine Hilfe für den Datenaustausch? *Mitteilungen und Berichte aus dem Institut für Museumskunde 31*, Berlin.
http://www.smb.museum/fileadmin/website/Institute/Institut_fuer_Museumsforschung/Publikationen/Mitteilungen/MIT031.pdf