

Text+ Consortium Progress Reports National Research Data Infrastructure (NFDI)

Progress Report Template Part 2, for internal use only

|  |  |
| --- | --- |
| Version | DRAFT |
| Versionsdatum | 2024-07-25 |
| Projekt | Text+ - Sprach- und textbasierte Forschungsdateninfrastruktur |
| Berichtszeitraum | 2023 |
| Grant Number DFG | *460033370* |

[1 Consortium 3](#_Toc172810486)

[1.1 Composition of the consortium and its embedding in the community of interest 4](#_Toc172810487)

[1.2 The consortium within the NFDI / General Contribution to the development of the NFDI 5](#_Toc172810488)

[1.3 International networking 6](#_Toc172810489)

[1.4 Organisational structure and viability / sustainability 8](#_Toc172810490)

[1.5 Operating model 9](#_Toc172810491)

[2 Research Data Management Strategy 10](#_Toc172810492)

[2.1 Scientific relevance and quality of the measures 12](#_Toc172810493)

[2.2 Metadata standards and reliable services 19](#_Toc172810494)

[2.3 Implementation of the FAIR principles and data quality assurance 23](#_Toc172810495)

[2.4 Services provided by the consortium 26](#_Toc172810496)

[2.5 Impact of changes of external conditions / constraints 28](#_Toc172810497)

[3 Additional Aspects 30](#_Toc172810498)

[4 Spending of Funds 31](#_Toc172810499)

[4.1 Description and Summary of Contributions by (Co-) Applicants 33](#_Toc172810500)

[4.2 Description of the cash flow developments 33](#_Toc172810501)

[1 List of outputs produced by the consortium 34](#_Toc172810502)

[2 Bibliography 34](#_Toc172810503)

[3 Additional comments on data sheet 34](#_Toc172810504)

[4 Data sheet 34](#_Toc172810505)

* 1. Consortium

|  |  |
| --- | --- |
| Institution | is also participating in NFDI consortium |
| Leibniz-Institut für Deutsche Sprache | KonsortSWD |
| Berlin-Brandenburgische Akademie der Wissenschaften | NFDI4Culture, NFDI4Objects |
| Deutsche Nationalbibliothek | NFDI4Culture, NFDI4Memory, NFDI4Objects, Base4NFDI |
| Georg-August-Universität Göttingen | NFDI4Microbiota, NFDI4Health, NFDI4Biodiversity, NFDI4Culture, NFDI4Memory, NFDI4Objects, NFDI4Chem, NFDI4BIOIMAGE, FAIRAgro, PUNCH4NFDI, NFDI4Earth, DAPHNE4NFDI, Base4NFDI |
| Akademie der Wissenschaften und der Literatur Mainz | NFDI4Culture, NFDI4Memory, NFDI4Objects |
| Albert-Ludwigs-Universität Freiburg | DataPLANT, NFDI4BIOIMAGE, NFDI4Energy, PUNCH4NFDI, NFDI-MatWerk |
| Bayerische Akademie der Wissenschaften | NFDI4Ing, NFDI4Culture, NFDI4Memory, NFDI4Objects, NFDIxCS, PUNCH4NFDI, NFDI4Earth, BERD@NFDI |
| Eberhard Karls Universität Tübingen | GHGA, DataPLANT, NFDI4Memory, NFDI4Objects, NFDI4Immuno, NFDI4Earth, DAPHNE4NFDI |
| Forschungszentrum Jülich | NFDI4Microbiota, NFDI4Ing, DataPLANT, NFDIxCS, NFDI4BIOIMAGE, FAIRAgro, PUNCH4NFDI, NFDI-MatWerk, NFDI4Earth, DAPHNE4NFDI |
| Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen | NFDI4Ing, NFDI4Biodiversity, NFDIxCS, KonsortSWD |
| Heidelberger Akademie der Wissenschaften | GHGA, NFDI4Objects |
| Herzog August Bibliothek Wolfenbüttel | NFDI4Memory |
| Hochschule Darmstadt | NFDI4Ing, NFDI4Memory, PUNCH4NFDI |
| Julius-Maximilians-Universität Würzburg | NFDI4Microbiota, NFDI4Chem |
| Klassik Stiftung Weimar | NFDI4Culture, NFDI4Objects |
| Ludwig-Maximilians-Universität München | NFDI4Microbiota, KonsortSWD, GHGA, DataPLANT, NFDI4Culture, NFDI4Memory, NFDIxCS, PUNCH4NFDI, NFDI4Earth, DAPHNE4NFDI, BERD@NFDI, MaRDI |
| Max Weber Stiftung - Deutsche Geisteswissenschaftliche Institute im Ausland | NFDI4Culture |
| Niedersächsische Akademie der Wissenschaften zu Göttingen | NFDI4Memory |
| Otto-Friedrich-Universität Bamberg | NFDI4Memory, NFDI4Objects, NFDIxCS |
| Technische Universität Darmstadt | NFDI4Microbiota, NFDI4Ing, NFDI4Objects, NFDIxCS, PUNCH4NFDI, NFDI-MatWerk, NFDI4Earth |
| Technische Universität Dresden | NFDI4Ing, GHGA, NFDI4Culture, NFDI4Chem, Base4NFDI, NFDI4Immuno, NFDIxCS, NFDI4BIOIMAGE, PUNCH4NFDI, NFDI4Earth, NFDI4DataScience |
| Universität Duisburg-Essen | NFDI4Microbiota, NFDI4Ing, NFDI4Health, NFDI4Biodiversity, KonsortSWD, NFDIxCS |
| Universität Hamburg | NFDI4Microbiota, NFDI4Memory, NFDI4Objects, NFDIxCS, PUNCH4NFDI, NFDI4Earth, NFDI4DataScience, BERD@NFDI |
| Universität Paderborn | NFDI4Ing, NFDI4Culture, NFDIxCS, NFDI-MatWerk |
| Universität Trier | NFDI4Culture, NFDI4Memory |
| Universität des Saarlandes | NFDI4Ing, GHGA, NFDIxCS, NFDI-MatWerk |
| Universität zu Köln | NFDI4Microbiota, NFDI4Health, GHGA, NFDI4Culture, NFDI4Memory, NFDI4Objects, NFDI4BIOIMAGE, PUNCH4NFDI, NFDI4Earth, BERD@NFDI |

* + 1. Composition of the consortium and its embedding in the community of interest

The consortium began with 30 funded partners and four contributing in kind. All original partners remain, and three new partners have joined:

* The (1) *University of Marburg*: The Research Centre Deutscher Sprachatlas, a data centre for linguistic data since 1876 joined and added expertise in Lexical Resources. It collaborates in training and development without project funding.
* The (2) *University of Wuppertal*andthe (3) *University of Applied Sciences in Wismar*: They integrated previous work into the Text+ infrastructure through cooperation projects and are committed to extending their contributions beyond the one-year funding period of the cooperation projects.

These additions have expanded Text+ in the fields of expertise, resources, and collaboration with the community, contributing to the consortium's objectives.

Cooperation projects have been an efficient dissemination tool for Text+ services, while also allowing community influence on the infrastructure provided. Calls for contributions were well-received by the community, resulting in a large number of applications. In the first two rounds, 14 projects were funded, with more under review in the third round. These projects continue to expand Text+.

Communities of interest play an active role in the Text+ governance through community-based panels. For the data domains (Collections, Lexical Resources, Editions), these are the Scientific Coordination Committees (SCCs), and for the task area Infrastructure/Operations, it is the Operations Coordination Committee (OCC). Initially these committees were filled by nominations from supporting learned societies, these committees now include members elected from outside of the consortium. The selection process of members ensures timely and efficient addressing of the community needs.

The consortium benefits communities by providing data, services, and knowledge sharing, integrating diverse research efforts, and developing new methodologies. Collaboration with other consortia, such as NFDI4Culture and KonsortSWD, and joint contributions to relevant conferences enhance these benefits. Text+ also participates in the cross-consortial initiative Base4NFDI, which aims at developing core services and methods for all consortia, contributing to the overall research system.

* + 1. The consortium within the NFDI / General Contribution to the development of the NFDI

Text+ is actively involved in the development and establishment of the National Research Data Infrastructure (NFDI) in Germany and is part of all structures established. As such Text+ participates in the monthly meetings of the NFDI spokesperson and management meetings, NFDI finance and communication meetings, which include representatives from all consortia. NFDI’s consortial meeting, was chaired by the consortium speaker from Text+ from 2022 to 2024. This highlights the role of Text+ within NFDI governance structure. Part of this engagement was the contributions to the mission and vision statement of the NFDI as a whole.

Text+ is part of all sections of NFDI and the cross-cutting topics discussed there and in dedicated task forces. The sections within the NFDI Association, include: Common Infrastructure (19 participants from within Text+); Ethical, Legal, and Social Aspects (6 Text+ participants); Industry Engagement (3 Text+ participants); (Meta)data, Terminologies, and Provenance (17 Text+ participants); Training and Education (9 Text+ participants). Additionally, Text+ is part of several NFDI task forces, such as those focused on evaluation and reporting (co-chairing this group), and governance.

The collaboration within the NFDI covers close interaction with other consortia. A full data sheet with all collaborations within NFDI between all consortia is available online (DOI: <https://doi.org/10.5281/zenodo.12819087>). One critical aspect is the collaboration of the four NFDI consortia from the humanities, for which a Memorandum of Understanding (DOI <https://dois.org/10.5281/zenodo.3265763>) provides insights in the modes of collaborations and delimiting the target disciplines of the consortia. The speakers of these consortia and the coordinators of the consortia meet regularly for interchange and common activities, such as information booths at conferences addressing cross-disciplinary communities. Another memorandum of understanding was created with KonsortSWD from the Social Sciences to collaborate with regards to reusable data sets and methods. Text+ organized a "Social Media Data" working group, in collaboration with NFDI4Culture, BERD@NFDI, and KonsortSWD, with a virtual workshop series titled "Show & Tell."

Text+ was contributing to the application process for Base4NFDI, contributing to working groups and organizing workshops to develop and refine the application concept, with a Co-spokesperson of Text+ also contributing to Base4NFDI in the same capacity. Basic services developed in Base4NFDI are closely monitored as they emerge and are prepared for integration, such as IAM4NFDI for the authorization and identification infrastructure, PID4NFDI for persistent identification etc.

Text+ has been instrumental in the NFDI's development, contributing to essential cross-cutting topics, and fostering collaboration across consortia.

* + 1. International networking

International activities were the starting points of Text+, with the national branches of the European Research Infrastructure Consortia (ERIC) CLARIN and DARIAH being foundational for the national activities of Text+. The consortium for Text+ was created with partners from these networks, along with additional institutions active nationally. Hence, from its beginning, Text+ has been dedicated not only to national but also international networks, making international networking a particular strength of Text+.

The internationalisation strategy of Text+ relies on three main components:

* Cooperation with European Research Infrastructure Consortia (ERICs) and other European cloud services catering to similar disciplines and user groups.
* Institutionalised international cooperation among partner institutions of Text+ in the field of Research Data Management (RDM).
* Participation of individual experts from within Text+ in international organisations for setting standards and developing RDM practices.

Text+ has close ties with ERICs such as CLARIN ERIC and DARIAH ERIC, both serving the European Social Sciences and Humanities (SSH) community from different perspectives. Nearly all national partner institutions of these ERICs are also part of Text+, which allows Text+ to closely follow and contribute to their developments. Text+ participates in their annual meetings, presenting German developments. In 2025, the Text+ plenary will be organised in conjunction with the DARIAH ERIC annual meeting in Germany. Data and developments outlined by Text+ are available for international partners, including technical interfaces to harvest metadata on datasets provided by Text+. Past collaborations within these ERICs have led to involvement with the Social Sciences and Humanities Open Cloud (SSHOC, <https://www.sshopencloud.eu/>) as part of the European Open Science Cloud (EOSC) initiative.

Beyond ERICs, Text+ benefits from the strong international networks of its partner institutions in academia. These networks include, amongst others (1) ALLEA (European Federation of Academies of Sciences and Humanities), (2) CENL (Conference of European National Librarians), (3) LIBER (Association of European Research Libraries), (4) IFLA (International Federation of Library Associations and Institutions) (5) EFNIL (European Federation of National Institutions for Language).

Text+ also collaborates with organisations setting RDM standards, such as the Research Data Alliance, Text Encoding Initiative (TEI), and the International Organization for Standardization (ISO). For example, the international Nestor network works on creating standards for data centres, including the Nestor certification for trustworthy repositories. Text+ experts actively participate in these networks, facilitating the exchange between Text+ and international initiatives. Through all these networks, Text+ engages in the exchange of RDM practices and standards. One example of such activities is the initiation of a dialogue between research infrastructures and the European National Librarians. Previously, these institutions operated separately. Text+ was key in organising the "National Libraries as Data Dialogue Forum" in autumn 2022. Before this forum, the ERICs were not in close contact with CENL. The forum, co-initiated by members of Text+ who were previously not part of any ERIC, brought these groups together for further activities. It initiated a dialogue aimed at providing additional options for opening national library data for research purposes in the humanities.

These international cooperations are further complemented by research projects with the participation of Text+ partners, also addressing RDM. More detailed information on our internationalisation activities can be found on the Text+ website at <https://www.text-plus.org/vernetzung/international/>.

* + 1. Organisational structure and viability / sustainability

The operational structure of Text+ successfully aligns with the governance structure outlined in the project proposal and is codified in a legal contract between the applicant institution and each partner:

1. **Governance Structure:** The governance structure was effectively implemented with the following entities:

The **Institutional Board**, termed *Leitungsgruppe* in the legal documentation, consists of the official representatives of the five co-applicant institutions and takes final decisions on long-term strategic and financial issues. This board meets annually, making efficient decisions via circulation if required, for example, for the final decision on cooperation projects funded through flexible funds.

The **Steering Group**, legally termed *Steuerungsgruppe*, consists of the two spokespersons[[1]](#footnote-1) and the co-spokespersons, the vice speakers. The steering group meets on a monthly basis.

The **Scientific Board,** comprising the chairs of the Scientific Coordination Committees and the Operational Coordination Committee, along with the spokesperson of the consortium. The co-spokespersons participate to represent their Task Areas. The scientific board meets 4 times a year.

The **Coordination Committees**, one for each task area in Text+ include elected members of the community, the respective co-spokesperson, plus one or two experts from the Task Areas (the latter non-voting). The Coordination Committees meet according to a roadmap based on the cooperation project review cycle and monitoring of the project achievements in the individual task areas.

As additional structures, each Task Areas Assembly, consisting of everyone involved in each task area, meets regularly. Depending on the size of the Task Areas, thematic subgroups have been formed for practical work. The Text+ Plenary, which meets once a year and includes all partners and researchers within Text+, invites the Coordination Committee members and the community. This structure follows the governance as laid out in the application, with extensions made to address diversity.

To support the steering group, a group was created to prepare documents and handle non-strategic coordination, with representatives from each Task Area leader.

2.  **Financial Plans and Fund Distribution:** Fund distribution adhered to the proposal’s financial plans and is fixed in the cooperation contract. This contract can be altered only by a unanimous partner vote. Flexible funds for cooperation projects were unspent in the first year due to the review process specification, transferring into the second year and doubling the available funds. The funds budgeted for cooperation projects were spent according to the established review procedure. However, budget cuts led to one partner teaming up with another to accomplish the intended workload, requiring a redistribution of funds and a partner vote. This process was initiated. Future contract amendments should simplify fund redistribution in such cases.

3. **Infrastructure Reliability and Sustainability:** The operational infrastructure established is highly reliable while allowing the flexibility to react to a changing environment. As such, additional overarching working groups were established beyond the initial application, covering: a working group on reference implementation for all data centres with their technical and organisational requirements; reaching out into the communities; website content creation; linked data; blog; registry; federated content search; cooperation with Specialized Information Service (FID) of libraries; research data management; and helpdesk.

Overall, the governance structure has proven to be stable and effective, and the financial plans have been successfully adhered to, ensuring the smooth operation and sustainability of the Text+ consortium.

* + 1. Operating model

Text+ has been recognized as highly valuable to the community. This is evidenced by an increasing number of projects in the application phase seeking advice on RDM questions, services, and collaboration. With 38 written statements from Text+ to support the projects and outline options where Text+ services can be used, the impact is considerable.

Our financial model and mode of operation are based on the direct involvement of partners and their contributions, along with funding granted by the NFDI program. This involvement results in clear responsibilities for all data and services. This model also ensures that services are tailored to the community's needs, as research-driven contributors to Text+ take responsibility according to their needs and those expressed by the community, for example, in the Coordination Committees. As research data management is emerging in the community, the community has provided feedback to require the services with minimal additional obstacles such as contracts and at no added costs.

Partners' contributions have been substantial, not only including financial support but also the provision of data and personnel crucial for the success of Text+. The computing and storage hardware, while not funded in the NFDI framework, has played a pivotal role in supporting our data storage and computing needs. This includes tasks requiring significant computational power. However, there are limits to this approach, particularly concerning the performance hardware required for advanced experiments with large language models (LLM) and concerning memory consumption. While current needs have been met through partners' contributions, scaling these efforts will likely require additional funds also for hardware.

In summary, the Text+ operation model has been effective in maintaining and developing our services. Continued collaboration will be essential to address future challenges and sustain the growth and impact of Text+.

* 1. Research Data Management Strategy

The research data management strategy of Text+ is oriented towards the three Grand Challenges outlined in the project proposal: (1) Mastering the Diversity of Research Data and the Diversity of Communities of Interest for Language and Text, (2) Assisting Humanities Scholars in the Re-use, Production, Preservation, and Innovative Use of Research Data, and (3) Develop a Governance Model that Supports Joint Responsibility and Consensus Building among Infrastructure Providers and Infrastructure Users as well as collaboration across disciplinary boundaries.

To address these challenges effectively, Text+ is organised into three distinct data domains, Collections, Lexical Resources and Editions. The work programs of these data-driven task areas, however, are organised in parallel around five common measures: Reference Implementation (M1), Portfolio Development (M2), Standardisation Activities (M3), Community Activities (M4) and Software Services (M5). They are complemented by the Infrastructure/Operations Task Area with its measures Findability (M1), Accessibility (M2), Interoperability and Re-Usability (M3), Community Activities (M4), and Software Services (M5).

Key for the successful implementation is the Text+ Data Space, establishing and improving findability, searchability and reusability of humanities research data. A data space, by definition[[2]](#footnote-2), integrates data from different sources by relying on existing mapping and matching techniques while leaving data sovereignty to the sources of the data. The Text+ Data Space consists of the Text+ Centres – data and competence centres providing RDM services – the Text+ Registry, the Federated Content Search (FCS) as well as the distributed services and tools provided by the participating Text+ Centres. The Text+ Centres create a distributed environment for existing data and data management of new data. By the distributed data management all aspects of complex legal states of the data, diversity of data types and disciplines can be addressed.

The Registry and the Federated Content Search (FCS) are two central services to provide easy access to the data. The Text+ Registry is a versatile and flexible system for describing and cataloguing diverse entities such as resources. The Federated Content Search is the central application in Text+ for searching directly in distributed research content. It supports searching in (often extensively annotated) data sets of Text+ Centres. The FCS is being developed in close cooperation with European partners and thus enables access to resources already available there and vice versa. The modular concept consisting of Text+ Centres is scalable and open for new partners as well as new data domains.

Quality assured Text+ Centres, for example by means of certifications, provide a stable environment for the community. The community also receives assistance by the non-technical services within Text+ consisting of consulting and community activities.
The services of this strand centre around the helpdesk and provide support a) in improving the quality of research data in line with the FAIR principles (Wilkinson *et al.*, 2016), b) to promote data literacy and RDM awareness especially for disciplines in transition towards data-driven research, c) for linking of the Text+ community to the overarching NFDI.

All services of Text+ are accessible via the Text+ portal (<https://www.text-plus.org>).

* + 1. Scientific relevance and quality of the measures

The **Text+ Data Space** has been successfully implemented in all data domains[[3]](#footnote-3). All data domains have been working intensively on the respective data and have implemented and maintained the **Text+ Centres** (CITE DATA Centre website). These data centres are key to the FAIR provision of the Text+ Data Space (see section 2.3). Distinguishing between the different data domains while maintaining a parallel structure of measures has proven very efficient in filling the Text+ Data Space adequately.

Yet, in order to counteract pillarisation processes, Text+ fosters strong collaboration between all task areas. This is evidenced by the expansion of joint working groups (see section 1.4), in which all data domains are represented, while at the same time working individually on the requirements of the different data. These working groups also extend the portfolio of Text+ by providing data management services (see <https://www.text-plus.org/daten-dienste/ingest/> ) to third parties (see also section 2.3).

To illustrate the success and scientific relevance, the following section focuses on the data domains’ measures M1 *Reference Implementation*, M2 *Portfolio Development* and M4 *Community activities* and the respective contributions of the data domains. Each one of them placed a systematic focus on these three measures in order to fulfill the primary purpose of Text+ – to build a joint portfolio of data, services and expertise to provide concrete solutions for the aforementioned Grand Challenges in the field of text- and language-based research data. M3 *Standardisation* will be covered in section 2.2 of this report and M5 *Software Services* in section 2.3.

Measure **M1 *Reference Implementation*** primarily aims at Grand Challenge 1 by establishing the Text+ Data Space. With regards to the Registry and considering the enormous heterogeneity of research data in the field of scholarly editing, TA Editions follows a distinct open approach. By cataloguing both internal and external edition projects, regardless of being provisioned by a Text+ Centre, the findability of print, hybrid, and (born-)digital editions is eminently improved (as desired in User Story [#410](https://text-plus.org/en/themen-dokumentation/user-storys-2020/user-story-410/)). Editions are recorded irrespective of their media form, adding value compared to existing catalogues of digital editions and library catalogues. TA Editions aims to represent the edition landscape as comprehensively as possible, without being limited to projects carried out in Germany or with German institutional participation. The registration process is conducted according to categories defined in a data model[[4]](#footnote-4) developed in constant consultation with community stakeholders, such as the SCC Editions – whose members represent various research associations in the field of scholarly editing – as well as external Text+ partners, particularly FIDs. This ensures the integration of the needs of a broad research community of scholarly editors and those from library and information studies perspectives. After successfully integrating the editions of all partners and information from the catalogue of AGATE and Gepris, TA Editions is focusing on developing a low-maintenance curation model and dissemination strategy to increase the number of entries and provides continuous support and community feedback options to improve the Editions Registry both qualitatively and quantitatively (see sections on Community Activities and Text+ Consulting below).

TA Lexical Resources provided a comprehensive account on current implementations of repositories at the member institutions[[5]](#footnote-5) in which technical and organisational procedures and constraints are described and compared showing broad similarities across all data centres. This finding is generally in line with the overall situation as surveyed and discussed by the project-wide working group Reference Implementation.

Building up a framework that allows scientists to browse many distributed repositories simultaneously with a single search query was the main focus of TA Collections (the lack of interlinked repositories is criticised in User Story [#316](https://text-plus.org/themen-dokumentation/user-storys-2020/user-story-316/), for example). The successfully established distributed infrastructure takes into account the diversity of the centres’ holdings as well as the different accessibility levels due to various legal restrictions[[6]](#footnote-6). This flexible approach reflects the willingness of TA Collections to expand its scope towards new types of data, understanding the term *collections* very broadly towards all text and language based research data used by the target audience.[[7]](#footnote-7)

In **M2 *Portfolio Development*,** Text+ established various procedures to increase its portfolio – starting with the extension and integration of the data held by the Text+ Centres (see the Text+ Registry[[8]](#footnote-8)), expanding on new data sets of third parties and including new partners with their respective holdings. A special highlight with regards to new partners are the holdings of the Deutscher Sprachatlas, already mentioned in section 1.1. Additional data sets by third parties were acquired by the help of cooperation projects and the Text+ community based governance.

To address Grand Challenge 3, each Task Area is accompanied by a Coordination Committee consisting of representatives of the respective research communities. The collaboration with the Coordination Committees aims, among other aspects, to expand the portfolio of data, services and expertise. In the course of the Text+ cooperation projects, the CCs acted as reviewers in the integration of 14 new projects[[9]](#footnote-9) and their associated data, services and expertise into the Text+ Data Space, the Text+ Consulting, Helpdesk and Community Activities. Collaboration with cooperation projects significantly fostered the development of the Text+ infrastructure as most of those projects broadened the diversity of the types of data to be handled by the infrastructure. Notable examples are the inclusion of georeferenced lexical data as well as the emerging support for resources in non-Latin scripts for the FCS in the TA Lexical Resources, the integration of substantial new corpora such as the digital program archive (DIPA) of the German adult education centres or the corpus of modernist literature (KOLIMO) in the TA Collections, and the integration of data and consulting expertise from cooperation projects Pessoa Digital, Edition2LD and Friver+ in TA Editions.

Beyond their function within the Text+ governance, the CC members have quickly proven to be cornerstones of community involvement by bringing in perspectives and needs of their respective communities, thus helping shape the work program of Text+ and strengthen the ties with their designating academic societies (see Section 1.1).

In the reporting period, the focus of the work in TA Lexical Resources was on providing unified and consolidated access to all lexical resources provided by Text+ (which is one of the demands of User Story [#501](https://text-plus.org/themen-dokumentation/user-storys-2020/user-story-501/)). Member institutions of the TA focused primarily on the integration of their own extensive data into the common framework of the FCS’s lexical query tier. By now, 27 dictionaries[[10]](#footnote-10) and lexicons are operational and can be queried by the FCS. Next, the focus will shift to the curation and integration of external resources. This will entail ongoing coordination with the activities related to M1. The selection of scientifically relevant external resources will again be supported and supervised by the SCC.

TA Collections contributed many newly developed endpoints to connect substantial text corpora to the FCS, such as the German Newspaper Portal (which is a good start for newspapers and magazines, but is not enough, see User Story [#309](https://text-plus.org/themen-dokumentation/user-storys-2020/user-story-309/))[[11]](#footnote-11). Future developments will enable the FCS to also report search hits in corpuses that are subject to legal restrictions by only stating that hits were found without showing actual text snippets. This progress will entail many more collections to be made searchable via the FCS.

In order to integrate research data beyond the centres’ present holdings into the Text+ infrastructure, Text+ provides depositing services[[12]](#footnote-12) (see User Story [#351](https://text-plus.org/themen-dokumentation/user-storys-2020/user-story-351/) for requirements for depositing; a need for a hosting service is also shown in User Story [#510](https://text-plus.org/themen-dokumentation/user-storys-2020/user-story-510/)) – which provides a workflow that helps scholars to decide which of the specialised Text+ centres is most appropriate for the research data that they want to be archived and made available by Text+[[13]](#footnote-13). An integral part of this service is to provide advice and support for enhancing data quality if needed. Text+’s data depositing service is disseminated through various community activities (see the section below).

In addition to the open approach regarding editions within the Text+ Data Space, the development of the TA Editions portfolio through cooperation projects was achieved on the one hand through the integration of new editions and participation in the practical testing and improvement of the Registry data model. On the other hand, the cooperation projects have contributed their expertise especially in the areas of Linked Open Data and science communication to the portfolio of TA Editions. All cooperation projects have made contributions to the Text+ Blog as a prominent tool for outreach and dissemination of Text+.

Aligned with Grand Challenge 2, Text+ provides valuable support to its target groups within the scope of **M4 *Consulting and Community Activities*** – including individual researchers and their associations in the field of text- and language-based research data. This assistance is primarily offered through the Text+ Consulting framework, which encompasses the Text+ Helpdesk[[14]](#footnote-14), bi-weekly open consultation hours ("Research Rendezvous")[[15]](#footnote-15), numerous event formats and workshops on research data management as well as guest lectures, keynotes and appearances (the need for training is formulated in User Story [#418](https://text-plus.org/en/themen-dokumentation/user-storys-2020/user-story-418/) and others). Conceptually, Text+ emphasises the integration and cooperation with relevant stakeholders such as research networks, associations and FIDs. As to dissemination in general, the Text+ Blog[[16]](#footnote-16) has established itself as a central organ of outreach, where Text+ reports on past, ongoing, and upcoming events, publications, calls for submissions, and highlights the data culture of its community.

Following the successful launch of a coordinated consulting service integrated into the **Text+ Helpdesk,** a standardised model for the documentation of consultation procedures was developed. The aim was to provide a documentation template for tracking consultations, to facilitate handovers and for analyses based on standardised information. In this regard, a survey was designed and conducted to map existing expertise and competencies among all members of Text+ with respect to their fields of research. The results of the mapping were presented in the form of a blog post.[[17]](#footnote-17)

The objective was set in view of the insights of the consultations held to date, according to which consulting requests in the context of the NFDI are oftentimes aimed at the area of research data management (RDM) indeed, but clearly exceed the area of assistance, for example, with the creation of data management plans or in the form of recommendations for specific formats or repositories.

Due to our **consulting framework**, Text+ is therefore able to tackle questions regarding all aspects of text- and language based research data – from preliminary planning and the actual implementation to the (usually) subsequent archiving. The depositing services are also integrated with the RDM support provided by the consortium[[18]](#footnote-18). Assessment of legal and ethical issues is also covered by Text+’s helpdesk agents’ expertise. The service is aimed at our community as a whole – regardless of status or discipline.

Consulting is increasingly performed in collaboration with the humanities partner consortia NFDI4Culture, NFDI4Memory, and NFDI4Objects. This approach ensures that the specialised and transdisciplinary consulting needs of the humanities are effectively addressed, leveraging shared expertise across consortia. In the ongoing consolidation processes of the NFDI under the guiding principle of "One NFDI", the helpdesks of various consortia are increasingly interlinking to pool expertise, to provide seamless support and to further standardisation of consulting processes in NFDI and beyond. In this regard, Text+ participates in the working group Helpdesk of the NFDI section Training & Education.

Furthermore, other NFDI consortia have the option to utilise the technical infrastructure of the ticket system which is operated by GWDG and SUB Göttingen. Concrete testing has been conducted with KonsortSWD. This further underscores the Text+ helpdesk's role in providing crucial support to the research community and facilitating the efficient handling of inquiries.

Additionally, dedicated **community activity** formats have been developed to promote the services provided by Text+, with a particular focus on the Text+ Data Space. These formats include hackathons and "Edit-a-thons" aimed at integrating metadata, tools, and services into the Text+ Registry or connecting new corpora and even whole data centres to the FCS.

In order to involve different scientific communities into its work, TA Collections organises community workshops on a yearly basis that are open to all interested scholars. While the first workshop (*Korpora in Text+: Kennenlernen und nachhaltig nutzen*[[19]](#footnote-19)) highlighted specific collections available and yet to be integrated in Text+, the second one (*Wohin damit? Storing and reusing my language data*[[20]](#footnote-20)) promoted TA Collections’ data depositing service sparking multiple co-operations between attending scholars and Text+ Centres. The most recent workshop (*Sprechen verstehen: KI und gesprochene Sprache*[[21]](#footnote-21)) centred around automatic transcription of language data using artificial intelligence (AI) tools, thus targeting a completely different audience than the previous two workshops.

Beyond these workshops TA Collections produces output to support scholars in their day-to-day challenges. For example, a decision guidance[[22]](#footnote-22) – a byproduct of a paper encompassing all recent development in German copyright law[[23]](#footnote-23) (User Story [#601](https://text-plus.org/themen-dokumentation/user-storys-2020/user-story-601/), for example, refers to copyright) – helps scholars decide whether they are allowed by copyright law to conduct text and data mining on a specific corpus or not. A series of blog posts[[24]](#footnote-24) invites the scientific community to exchange about improvement of material-specific aspects in the DFG guidelines on digitization.

TA Editions acts as an open platform and multiplier that relies on the participation of all actors in the field of scholarly editions. To facilitate participation and integration of its target group into the complex of Text+ and NFDI, an Onboarding Guide has recently been published.[[25]](#footnote-25)

TA Editions is pursuing three primary goals: (1) networking and cooperating with stakeholders, (2) the expansion, improvement, and dissemination of Text+ services for editions, and (3) the improvement of awareness for sustainability and standardisation in compliance with the FAIR principles. To achieve this, TA Editions is involved in over 60 relevant annual events, either as an organiser or participant. In this context, TA Editions cooperates with FIDs, research associations and other consortia involved in the field of scholarly editing (e.g. several cross-consortial consulting cases, and co-applicant of basic service “nfdi.software” with NFDI4Culture, Community Cluster 9 “Objekte als Inschriftenträger” in NFDI4Objects). Due to the intensive collaboration with FIDs in the context of Registry and Text+ Consulting, a dedicated focus group with a regular exchange format, the FID/Text+ Jour Fixe, has been established. Prominent examples serving to specifically promote and provide hands-on support for creating and optimising entries in the Editions Registry is the “Edit-a-thon”, whose conceptual preparatory framework was realised within the context of the cooperation project Pessoa digital.

Regarding the third goal, TA Editions is developing guidelines for quality assessment and assurance through an open participation process, serving as a manual for the creation, editing, and publication of research data in the field of digital editions. The community activities are complemented by workshops, tutorials, and curricular recommendations on editing techniques and practices, as well as relevant tools. Scheduling and programmatic alignment are largely based on the community needs collected through Text+ Consulting. Notably, the annual FAIR February[[26]](#footnote-26) stands out, where Text+ engages to discuss the FAIR principles and their application in the context of scholarly editions.

* + 1. Metadata standards and reliable services

Text+ builds on existing standards and infrastructures and brings in established preparatory work and experience – notably when it comes to the Registry and the FCS. Among the standards used at the various data centres are: Dublin Core, DataCite, MARC21, ISO 24622-1 and ISO 24622-2 (CMDI), ISO 24619 (Persistent Identification of Language Resources), DCAT, LIDO, etc. The diversity of standards used within Text+ is very similar to the variety within the whole NFDI. Text+ therefore does not primarily focus on developing new metadata standards, but participates in both, the standardisation efforts within the NFDI as a whole and by providing interfaces for high quality metadata. By resorting to the participation of the data centres, Text+ ensures resource descriptions and metadata of high quality leading to a considerable depth of the Text+ search space. On the other hand, the combined expertise of the data centres provides for compliance with data domain-specific standards and practices, enforced by the subject-specific exchange within the thematic clusters of each data domain.

To emphasise findability, accessibility and reusability of the holdings of its centres and of its services, Text+ focuses on comprehensive descriptions of the various types of data and services in order to make them easily searchable for users. In Text+, the metadata schemas used for cataloguing the resources within central services such as the Text+ Registry, are coordinated first within the data centres of each data domain and second among the data domains.

The data model of TA Editions focuses on interoperability by integrating controlled vocabularies (see section 2.4). It is also expanded by an additional module to register edition tools and software via a Curated Tool Platform[[27]](#footnote-27) (tools are requested in User Story [#418](https://text-plus.org/en/themen-dokumentation/user-storys-2020/user-story-418/), for example), which is interoperable with the SSH Open Marketplace (<https://marketplace.sshopencloud.eu/>) and upcoming NFDI-wide solutions for Software Registries in the context of nfdi.software as upcoming basic service.

To foster the integration of metadata from existing repositories into the Text+ registry, TA Lexical Resources developed a common metadata model that is currently being deployed. This ensures a unified way of describing different types of lexical resources and making them findable across local infrastructures as well as within the Text+ Data Space.
TA Collections also devised a metadata schema[[28]](#footnote-28) in order to describe all kinds of collections that Text+ Centres may possess in a uniform manner. It respects the specificities of text and language data collections and yet shares many common properties with the description schemas of the other data domains, permitting generic descriptions of all resources available in the Text+ infrastructure.

To allow domain independent access, a working group harmonised the schemas, identifying common structures that can be used for cross domain search and that can be used also in the context of other NFDI consortia. To form a comprehensive catalogue, the Text+ Registry, corresponding fields of the three metadata schemas have been identified, tapping on the DataCite Metadata Properties[[29]](#footnote-29). These properties are resource and domain independent and thus allow standardised and uniform descriptions of any resources listed in the Text+ Registry. All additional resource-specific metadata provided by the data centres are displayed on the detailed description page of every resource – and are part of the search index.

The main objective has been to develop a concept that supports the diversity of the data domains and their resource types. A sophisticated approach has been implemented that facilitates the definition and evolution of the different data models e.g. to support additional entities. This way, resources exhibiting new properties that have not yet been catered for in the metadata description can be thoroughly modelled while maintaining compatibility with existing metadata descriptions. In addition, an import of data from existing sources into the Registry and manual editing and curation of resources have been designed to complement each other in order to be able to reuse existing information wherever possible and to fill gaps in a qualitative way by respective domain experts.

The Text+ Registry allows users to run complex search queries to find very specific resources as well as to freely explore, browse and filter search results by categories provided by the metadata schemas. In addition, the resource descriptions use identifiers from authority files and other reference sets for entity linking such as GND, Wikidata or DBpedia, e.g. for languages, creators, publishers, funders or subject headings. The data categories of the metadata are defined based on the given standards and established practices such as schema.org. That way, the Text+ Registry can interlink the listed resources and establish new and broader contexts for them. Currently, the Text+ registry comprises XXX records for YYY. It can be freely accessed at <https://registry.text-plus.org>.

In order to bundle standardisation efforts with other disciplines, Text+ works within the NFDI Metadata and Terminology Section which aims at defining common recommendations and standards for the whole of NFDI based on existing recommendations and standards such as DataCite and DCAT.

An important catalyst for interoperability of research data is the use of authority files. The use of authority files also eminently improves the interoperability of Text+’s holdings. A key element in Text+’s efforts to use and make available authority files is the creation and operation of the GND agency, contributing to the Integrated Authority File (GND). A significant difference of the GND from other authority files and sources for entity linking (e.g. Wikidata or DBPedia) important for research is its reliability of content and networking. The main goals of the Text+ GND Agency[[30]](#footnote-30), as one of the central services in the Text+ data space, is to ensure the semantically reliable interconnection of data and to promote the GND as an authority file in the research data domains of the Text+ community. Research data and metadata are enriched using methods of entity linking to the GND. For instance, editions or collections of unstructured texts, when linked to entities, significantly improve reusability for other researchers. This enhancement also facilitates linking resources at more granular levels beyond the entire dataset, promoting greater interconnectivity and interoperability across various resources and metadata. Additionally, expanding the usage of the GND can be achieved by encouraging and training researchers to contribute. This could involve proposing new entities or actively participating in the GND community, further enriching the database and its utility.

First and foremost the Text+ data domains are the primary target group and testbed for these activities. The GND Agency task in Text+ is a cross-cutting topic as it not only involves all data domains but also relies heavily on the existing GND community and – most importantly – on the cooperation with other GND agencies.

Other areas of standardisation are also on the meta-level of the underlying primary data, e.g. data formats utilised by Text+ that offer great potential for the research use of legally restricted resources, so called derived text formats (DTF)[[31]](#footnote-31). DTFs are usually created by enriching a given text corpus with different annotations and subsequently reducing information (e.g. by replacing tokens by their part-of-speech representation) until the resulting text format is no longer copyright relevant while still useful for certain research methods. Thus DTFs represent another way of making legally restricted resources available for the research community. Besides working on a new work item proposal for a DIN standard focussing on the technical creation of DTFs, TA Collections will also release a report on their legal implications. The publication of selected collections in specific DTFs by Text+ is planned in the near future.

Another data format is available for lexical resources. In the TA Lexical Resources, most member institutions maintain and provide their data in TEI based formats with a general tendency towards the adoption of the guidelines provided by the TEI Lex-0 initiative[[32]](#footnote-32) specifically. Standardisation processes both in TEI Lex-0 as well as in the general TEI are very active and are supported by their focus communities world-wide. Text+ members are actively participating in these processes, e.g. by participating in the TEI conferences, in TEI special interest groups and in regular meetings of the TEI Lex-0 initiative. They thus bring important impulses and requirements of Text+ into the international standardisation process.

With regard to standardisation activities, Text+ has been instrumental in establishing a working relationship between DIN and NFDI, with the signing of a Letter of Cooperation planned for October 2024. Text+ is also heavily involved in the DIN 105-00-06 AA "Language Resources and Language Technology" working group and ISO TC 37 SC 4 "Language Resources Management". Infrastructure/Operations is part of the DIN Strategy Forum on Digitisation (FOCUS group) and Text+ was involved in the standardisation roadmap AI (Wahlster and Winterhalter, 2022) Partners of Text+ are involved in the advancement of TEI, and there have been first tests for transferring lexical resources to ONTOLEX.

* + 1. Implementation of the FAIR principles and data quality assurance

Text+ is dedicated to the implementation of the FAIR principles. These guidelines will be disseminated in our communities through two fields of action: the services around the Text+ Data Space, developed within Text+ as well as contributed by the participating institutions, and through support and networking via Text+ Consulting, Text+ Helpdesk and Text+ Community Activities.

The key building blocks of the Text+ Data Space are the distributed Text+ Centres among the partner institutions which provide not only data itself but contribute a great experience in creating, providing and maintaining text and language-related research data. Corresponding to Grand Challenge I, our data is provided in different formats, with different annotations and analysed by different methods. In addition, it touches on the rights of third parties, e.g. through the copyright of texts and personal rights of spoken language recordings.

To address the diversity challenge with regards to Accessibility and Reusability, each data centre has agreed access and use restrictions with the right holders and makes the data available as openly as possible under the available licences. As these are usually non-transferable contractual documents between the partner institutions hosting the data and the rights holders, this means that the Text+ Data Space cannot exist in the form of a centralised single repository or long-term archive. Thus, Data centres must maintain their repositories and provide a solution for long-term archiving. The Text+ centres aim at documenting their reliability by means of certificates, either from Core Trust Seal (CTS, see https://www.coretrustseal.org/) or NESTOR (based on the standard DIN 31644, see<https://www.langzeitarchivierung.de/>).

Further key factors for addressing the diversity of data within the conceptual framework of the Text+ Data Space are (a) to provide users with information on existing datasets, available licences and access restrictions, (b) to explore ways of making data available within the limits set by legal obligations, and (c) to operate and maintain the interfaces between the central infrastructure components of Text+ and the distributed network of data centres.

(a) Information about the individual datasets within the Text+ Data Space is provided in the form of rich metadata, including persistent identifiers according to DIN ISO 24619:2018, such as a URN, handle, or DOI. Regarding the requirements of Findability, all information about datasets of the Text+ Data Space is accessible via the Text+ Registry (<https://registry.text-plus.org>), serving as a central discovery hub for information objects. It provides standard sets of metadata within a central and consistent backend for providing search options to users and employs specific data models to describe resources across the three data domains. Additionally, a fourth entity for services provided by Text+ partners as well as the open approach regarding the data domain Editions (see section 2.1) signifies the Registry's potential as in terms of flexibility, openness and adaptability to cater to entities beyond Collections, Editions and Lexical Resources. At the time of reporting, the Registry contains XXXX records and is being regularly expanded by testing additional entities and similar tools and approaches within the context of NFDI (e.g. AWK NRW as co-applicant of nfdi.software) and beyond (e.g. SSH Open Marketplace) to improve Interoperability.

As the usability of a backend as default view of Text+ Registry is not ideal, several options such as multi-faceted search, filtering or even generative LLM-based dialogue systems are currently under discussion.

(b) Text+ provides three options for access to the data in the data centres. (1) If rights permit, the data centres provide full access to their object data. As some datasets can only be provided with restrictions, data centres are encouraged to implement the IAM services (Identify and Access Management) being developed by Base4NFDI. This includes the Shibboleth services already in operation within the German Research Network (DFN, see [https://doku.tid.dfn.de/de:shibsp](https://doku.tid.dfn.de/de%3Ashibsp)). (2) For data stored in the data centres, Text+ provides the Text+ Federated Content Search (<https://fcs.text-plus.org/>). Building on top of previous developments, it allows a query to be sent in parallel to the federated data centres and the search results to be aggregated, depending on the availability constraints. Text+ has extended earlier work on FCS to include a search option for lexical resources (LexFCS[[33]](#footnote-33)). The thematic focus has been on expanding the amount of resources available, extending the FCS specification for new resource types, strengthening the overall software ecosystem and creating user-friendly web interfaces for different user groups. As a result, the FCS is now directly available via the Text+ portal and also forms the technical basis for the *decentralised dictionary platform* of the data domain Lexical Resources. Within the last three years, the number of resources that are available via the FCS has increased by over 130%. Prominent examples of newly integrated resources include the German Newspaper Portal of the German Digital Library, the German-language wordnet GermaNet and resources from the Deutsche Sprachatlas. Current and future work – in addition to continuing the above-mentioned topics – will in particular enable access to restricted resources via an IAM integration and the support of federated queries based on authority file records (e.g. using GND entities). The latter allows the disambiguated formulation of user queries and is a powerful tool to support semantic interlinkage between different resource types and languages.

FCS search results are also a special form of (3) DTF which, as described in section 2.2, reduce the information content of a dataset so that it can be used without liability restrictions. Data centres can make copyright protected resources in their repositories searchable via the FCS by limiting the amount of contextual words around the actual search hit until the resulting string is no longer copyright relevant.

(c) In addition to the provision of data already held in the data centres, the data centres also provide depositing services for researchers who are not (yet) members of the Text+ consortium and who collect and create data in the course of their research. There are several publications that describe the specialisations of the data centres[[34]](#footnote-34), which data can be deposited, and quality guidelines[[35]](#footnote-35) give orientation concerning data standards that should be met.

Based on the data centre profiles, users and data centres are matched up. This is currently part of the outreach process and may be supported in the future by automatic filtering or tool support. Once ingested, data provided to a Text+ Centre automatically becomes part of the Text+ Data Space, accessible via the Text+ Registry.

To improve the user experience and to communicate common processes, these have been streamlined to be included in communications with users, who are not familiar with standard depositing strategies such as those outlined in the OAIS standard (ISO 14721:2012). Common processes include the collection of metadata as required by the data centre, the determination of licences, the initial review of data for quality assurance and to check that the data matches the data centre’s profile, and internal curation and ingest procedures by the data centre without intervention from data providers (see <https://www.text-plus.org/daten-dienste/ingest/> ). During the ingest process, the entities in the metadata are linked to authority files, in particular to the GND.

A testbed for the successful implementation of data ingest and portfolio development processes are the cooperation projects, which are created to enable the integration of resources and services from outside of the consortium into the Text+ portfolio. With the defined evaluation, the first round of projects was completed by the end of 2023 with 10 projects that were initiated for the first two calendar years of the consortium, and a further 4 projects currently operational in 2024. For 2025, 8 applications were received, of which 4 were approved. The Coordination Committees consider the process and the collaborative projects to be successful, with the data repositories complying with the FAIR principles, which are also respected in the collaborative projects.

For an account of the contributions of Text+ Consulting and Text+ Comunity Acitivities to Data Quality and FAIR, see 2.1.

* + 1. Services provided by the consortium

Since 2023, Text+ has been utilising an architectural concept[[36]](#footnote-36) to depict its structure and to develop service offers tailored to basic use cases. These services can be grouped under two general headlines:

1. Making research data searchable and findable, accessible, re-usable (FAIR), and
2. providing the means to (long-term) archive research data, also according to the FAIR principles.

Both sets of services aim at creating a unified data space within Text+. While the first set comprises integrating and aggregating services that are built and maintained by Text+, the second set lies directly in the responsibility of the participating data centres[[37]](#footnote-37).

Resource specific services such as query endpoints for specific corpora or dictionaries or autonomous online presentations of editions, dictionaries or corpora are generally provided and sustained by each individual Text+ centre in the context of independently managed and financed projects. Especially the content related development of resources remains outside the scope of Text+. Collaboration with external projects on the level of content creation is restricted to consulting and technical advice; data sovereignty for collaborating projects is assured at all times. Technical services created and maintained within the Text+ project focus exclusively on overarching features for every individual resource (such as metadata aggregation in the Text+ Registry to foster findability), or across a broad range of resources (such as content aggregation in the Text+ Federated Content Search, an important example of applied interoperability). Therefore, a clear responsibility distribution between Text+ and collaborating projects or data providers is established: specific resources and interfaces to those resources are maintained by and within collaborating projects, while the unified and coherent dataspace that integrates those resources along the FAIR principles is maintained by Text+.

All services are accessible through the portal which acts as a user-friendly gateway. A good starting point to browse the portfolio is the Text+ Registry and the service descriptions of the contributing data centres. However, any demand by users for consultation, e.g. in case they don’t find a suitable resource or service, may be addressed by the helpdesk.

A comprehensive list of all services provided by Text+ can be found on the project website.[[38]](#footnote-38) In the remainder of this section, we will highlight a few examples.

There has been an increase in the use of authority files for entity linking across all TAs, for example using the Integrated Authority File (“Gemeinsame Normdatei”, GND). To fill gaps and maintain a disciplinary point of reference and to ultimately foster interoperability, Text+ established the Text+ GND Agency as a service and point of contact for the community (see section 2.2 for details).

A further example for data integration and applied interoperability are ready-to-use software services developed specifically within TA Collections: Natural Language Processing (NLP) pipelines and tools such as MONAPipe[[39]](#footnote-39) and Weblicht[[40]](#footnote-40) are easily accessible and user-friendly text processing services. They offer current text and data mining (TDM) and annotation features. Besides, they are open to new community-developed components, some of which were already successfully integrated, e.g. an *Entity Recognizer* detecting named entities and nous referencing literary characters[[41]](#footnote-41).

The inclusion of significant quantities of lexical resources in the Text+ infrastructure resulted in the adaptation and extension of the FCS to support queries on hierarchically structured lexical data as opposed to the querying of sequence-based annotation layers of running text. This was developed in a user driven process. Together with the currently funded collaboration project “Thesaurus Linguae Aegyptiae – More Fair with APIs”, work will continue to support non-Latin scripts in all services provided for lexical resources.

Guided by TA Editions, a key contribution of Text+ addressing a community of scholarly editors is the open registry for editions as part of the Text+ Registry, which significantly enhances the discoverability of printed, hybrid, and born-digital editions.[[42]](#footnote-42) In this regard, the Text+ Registry goes beyond cataloguing the portfolio of Text+ centres and establishes an interoperable, open search space for editions (and the software used) through the inclusion of relevant catalogues and the curatorially guided integration of as many editions as possible. By incorporating controlled vocabularies, it can also express relationships between the entities.[[43]](#footnote-43)

Another essential service for the community of researchers, teachers, and learners in the field of scholarly editions is provided by Text+ Consulting. The Text+ Helpdesk serves as an entry point to a consulting network that includes representatives from relevant research associations, specialist information services, and further stakeholders. In close collaboration with other initiatives in the context of the NFDI and beyond, we aim to improve data quality and standardisation possibilities in the field of scholarly editions.

* + 1. Impact of changes of external conditions / constraints

Since its inception, the Text+ consortium has witnessed several significant developments that have influenced the infrastructure landscape, services, and community needs within our field. Notable examples that will be discussed in the following are (1) technological advancements, such as the widespread availability of large language models (LLMs), (2) structural changes within the NFDI, especially the establishment of Base4NFDI, a cross-consortial initiative.

One of the most notable technological advancements during the reported period was the general availability of LLMs as disruptive technology, exemplified by the release of ChatGPT in late 2022. Although the use of language models is not new, and with previous datasets on embeddings, language models etc. well known in the consortium, the broad accessibility of large and generative language models introduces new opportunities and challenges for research, infrastructure, and the Text+ consortium. This development was not anticipated when the initial application was submitted and will take considerable time to fully integrate—if at all. Text+ is now focusing on several interrelated questions.

The first critical question is which research tasks can benefit from LLMs and how Text+ services can be enhanced using LLM and RAG (Retrieval-Augmented Generation) technologies. Secondly, how can language models be optimized for specific tasks? Thirdly, how can data from Text+ be integrated into large language models directly or via RAG technologies? Legal constraints, such as licensing agreements for original data provided by publishers and copyright laws, present obstacles to publishing language models created or tuned by Text+ partners, despite machine learning exceptions in national copyright legislation. One way to address this is by focusing on derived text formats and assessing their suitability for training or enriching language models.

Text+ has formed a cross-task-area working group to address these issues and has begun developing a technical platform for LLM-based solutions. Additionally, Text+ is collaborating with the new DHd working group "Applied Generative AI in the Digital Humanities" and European initiatives like the European Data Language Space (LDS) to support the development of European language models. Lastly, the 3rd Text+ Plenary in October 2024 will focus entirely on LLMs.

LLM and the discussion of the legal implications are also good examples for a changed legal background, which is essential for the work of Text+. Hence, Text+ closely monitors and influences developments together with the NFDI section on legal issues. This includes for example the AI Act, the EU Data Act, and the Research Data Act. Text+ experts have played a key role in drafting NFDI statements on legal aspects, including *Stellungnahme zum Daten-Governance-Gesetz* (published in May 2023), *Stellungnahme zur öffentlichen Konsultation zum Forschungsdatengesetz* (published in May 2023), *Comments on the draft implementing regulation concerning the availability of public data sets* (published in June 2022), and *Comments on the EU Data Act* (published in May 2022).

The most significant structural change within the NFDI was in the third NFDI call for proposals, starting on March 1, 2023, which included the Base4NFDI initiative aimed at providing and establishing basic services for the NFDI. Text+ actively participates in developing these basic services, specifically in applicant role in IAM4NFDI (authentication infrastructure), PID4NFDI (persistent identifier), Jupyter4NFDI (JupyterHub for the NFDI) and nfdi.software. Additionally, Text+ participates in the other applications, particularly in the requirements analysis (initialisation phase) for basic services, representing the needs of the specialist communities within Text+, and in the integration phase, contributing to the day-to-day work of the consortium and beyond. The service stewards for “federated storage- and compute-infrastructure” and “semantic interoperability and reusability” within Base4NFDI are collocated with a Text+ partner. Finally, Text+ also provides one of the co-spokespersons for Base4NFDI in the Governance Task Area.

* 1. Additional Aspects

Text+ is committed to promoting diversity across disciplines, gender, academic interests, and career levels. Initially, the voting members of the steering group consisted of 4 men and 1 woman, after reorganisation at one co-applicant institute, all voting members were male. Recognizing the imbalance in the gender ratio, also in the number of other female principal investigators (PIs) named by partner institutions, Text+ implemented several measures:

1. **Scientific Board Composition:** To counterbalance the predominantly male steering board, special emphasis was taken for the composition of the scientific board. As a result it turned out that the voting members, the chairpersons of the coordination committees, are predominantly female.
2. **Election Process for the Coordination Committees:** To ensure a balanced approach, the election process for coordination committees explicitly references gender balance. The election board ensures a balanced number of candidates, inviting additional nominees if necessary. The initial election resulted in female representation of 50–60 % in the Scientific Coordination Committees and 30 % in the Operation Coordination Committee.
3. **Steering Board:** the governance structure stipulates that voting members are the co-spokesperson and consortium spokesperson, these are appointed by the (co-)applicant institutions. This led to a male dominance of the board. To address diversity, the composition of the steering board was adjusted and the vice speakers were added to the steering board. This change demonstrates a commitment to gender awareness. It is intended to change the cooperation contract of all partner institutions in a way that female representation among the voting members of the steering board is assured, even if contractual changes pose a significant obstacle involving all legal partner institutions.
4. **Events:** For each event, such as the Text+ plenary, programme committees are asked to explicitly take a gender-balanced approach. This has been successfully achieved in past events, excluding presentations by the (co-)spokespersons.

Currently the proportion of women among the employees in the project is 37 % (as of the status report of the partners for December 2023). This is a major achievement, especially as recruiting women in technical domains is still challenging given the current situation on the labour market.

**Career Level Support:** Text+ supports researchers at all career levels. Research Data Management, which demands high expertise, involves a substantial number of postdoctoral researchers (26 %) and doctoral researchers (10 %). Notably, 45 % of researchers in Text+ are pre-doctoral. Additional staff includes professors and supporting staff.

**Disciplinary Diversity:** Researchers in Text+ come from a wide range of backgrounds, reflecting the *DFG Classification of Scientific Disciplines, Research Areas, Review Boards, and Subject Areas (2024–2028)*: Greek and Latin Philology; General and Comparative Linguistics; Experimental Linguistics, Typology, Non-European Languages; Individual Linguistics, Historical Linguistics; Applied Linguistics, Computational Linguistics; German Medieval Studies (Medieval German Literature); German Literary and Cultural Studies (Modern German Literature); European and American Literary and Cultural Studies; General and Comparative Literary Studies; Cultural Studies; History of Philosophy; Theoretical Philosophy; Practical Philosophy; Political Science; Musicology; Computer Science.

This diversity ensures that Text+ effectively addresses the needs of humanities disciplines working with text and language-related data.

* 1. Spending of Funds

Overall Spending by Institution

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **year 1****(Oct-Dec)** | **year 2** | **year 3** | **year 4** | **year 5** | **year 6****(Jan-Sep)** | **Total****in €** |
| **Institution** | **Totals in €** |
| **IDS** |  |  |  |  |  |  |  |
| **BBAW** |  |  |  |  |  |  |  |
| **DNB** |  |  |  |  |  |  |  |
| **NRWAW** |  |  |  |  |  |  |  |
| **SUB** |  |  |  |  |  |  |  |
| **AdWHH** |  |  |  |  |  |  |  |
| **AdWMZ** |  |  |  |  |  |  |  |
| **TUDA** |  |  |  |  |  |  |  |
| **USLDA** |  |  |  |  |  |  |  |
| **h\_DA** |  |  |  |  |  |  |  |
| **GWDG** |  |  |  |  |  |  |  |
| **HAB** |  |  |  |  |  |  |  |
| **JSC** |  |  |  |  |  |  |  |
| **Leopoldina** |  |  |  |  |  |  |  |
| **LMU** |  |  |  |  |  |  |  |
| **MWS** |  |  |  |  |  |  |  |
| **SAW** |  |  |  |  |  |  |  |
| **SLUni** |  |  |  |  |  |  |  |
| **STI** |  |  |  |  |  |  |  |
| **TUDD** |  |  |  |  |  |  |  |
| **UniBA** |  |  |  |  |  |  |  |
| **UniDUE** |  |  |  |  |  |  |  |
| **UniFR** |  |  |  |  |  |  |  |
| **UniHH** |  |  |  |  |  |  |  |
| **UniK** |  |  |  |  |  |  |  |
| **UniPB** |  |  |  |  |  |  |  |
| **UniTR** |  |  |  |  |  |  |  |
| **UniTÜ** |  |  |  |  |  |  |  |
| **UniWÜ** |  |  |  |  |  |  |  |
| **Total spending per year** |  |  |  |  |  |  |  |

* + 1. Description and Summary of Contributions by (Co-) Applicants

All partners contribute significantly in-kind by providing workspaces with appropriate technical infrastructure and related resources. They also contribute to and maintain existing datasets. Additionally, all partners have allocated extra staff to support the consortium's goals and endeavors. This commitment is evident as the 32 full-time equivalents (FTE) funded by the consortium are supplemented by 23 FTE provided in-kind by consortium members. All in all this means that approximately 72% of the project funding is matched by contributions from the partners.

* + 1. Description of the cash flow developments

According to the DFG's approval letter the reduction ratio of the total budget requested is 29.81%, spread evenly over the years. The DFG did not specify how exactly the cuts should be implemented. Different strategies were discussed. The final implementation was based on the following principles: 1. No partner was supposed to be left out. 2. In general small partners should not be reduced below viability, and 3. Central tasks should not be reduced if possible. Hence specific milestones and deliverables were reduced, in some implementations, the scope of the data to be analysed was reduced to enable keeping the principle structure. The cuts are characterised by a progressive procedure in which the largest reductions of 32% is imposed on the major contract partners with an original requested funding of over €1 million and. Based on this, the cut rates were linearly decreased for all other partners in proportion to the funding amount with a minimum cut rate of 27.13%. The administrative position (Measure 1 in the Administration task area) and the central funds for expenses, which are allocated in the Administration task area, are excluded from the reductions.

Appendix

The appendix may only include the following information and documents:

* 1. List of outputs produced by the consortium

Please list the different kinds of outputs (in the broad sense of the classes of the FRBR-aligned Bibliographic Ontology, FaBiO[[44]](#footnote-44)) that you have produced and published briefly, but including at least the following data: title, year, and persistent identifier / web link.

* 1. Bibliography

Wahlster, W. and Winterhalter, C. (2022) *Deutsche Normungsroadmap Künstliche Intelligenz – Ausgabe 2*. DIN and DKE. Available at: www.din.de/go/normungsroadmapki.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., ’t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016) ‘The FAIR Guiding Principles for scientific data management and stewardship’, *Scientific Data*, 3(1), p. 160018. Available at: https://doi.org/10.1038/sdata.2016.18.

* 1. Additional comments on data sheet

You can include an optional appendix (max. three pages) with additional comments on your entries in the data sheet (appendix 4).

* 1. Data sheet

It is mandatory to fill out the *Supplementary data sheet for consortia of the National Research Data Infrastructure (NFDI)* (DFG form nfdi1000). Please upload a PDF copy of the completed data sheet via elan as a separate document and additionally send the original document (Excel file) via e-mail to nfdi@dfg.de.

1. Unlike other consortia in the NFDI, Text+ follows the dual structure of NFDI by having one spokesperson in the consortium according to the NFDI association, a legal entity according to German law, and one according to the funded project. This was also an effective means to prepare for the takeover of the spokesperson role by the designated new spokesperson before the change took effect. [↑](#footnote-ref-1)
2. see <https://en.wikipedia.org/wiki/Dataspaces> [↑](#footnote-ref-2)
3. see Collections’ report <https://zenodo.org/doi/10.5281/zenodo.12771255> and deliverables LR1.2, E1.1, E1.2, E1.3 [↑](#footnote-ref-3)
4. see section 2.3 for details of the Registry data models for each data domain [↑](#footnote-ref-4)
5. see deliverable LR1.1 [↑](#footnote-ref-5)
6. see Collections’ report <https://zenodo.org/doi/10.5281/zenodo.12759059> [↑](#footnote-ref-6)
7. On the Text+ website we describe example data types that will be part of the TA Collections portfolio {Link} [↑](#footnote-ref-7)
8. See <https://registry.text-plus.org/>. [↑](#footnote-ref-8)
9. A complete list of the cooperation projects can be found here: <https://text-plus.org/vernetzung/kooperationsprojekte/#gef%C3%B6rderte-kooperationsprojekte-der-2-ausschreibungsrunde-2023> [↑](#footnote-ref-9)
10. See <https://fcs.text-plus.org/?&queryType=lex> [↑](#footnote-ref-10)
11. See <https://www.deutsche-digitale-bibliothek.de/newspaper> [↑](#footnote-ref-11)
12. See <https://www.text-plus.org/daten-dienste/ingest/> [↑](#footnote-ref-12)
13. Guidelines for integrating collections into Text+ are outlined here: [https://zenodo.org/doi/10.5281/ zenodo.12744055](https://zenodo.org/doi/10.5281/zenodo.12744055) [↑](#footnote-ref-13)
14. see <https://text-plus.org/helpdesk/> [↑](#footnote-ref-14)
15. see <https://events.gwdg.de/category/208/> [↑](#footnote-ref-15)
16. Link: <https://textplus.hypotheses.org/>. [↑](#footnote-ref-16)
17. Daniela Schulz: Was kannst Du, was ich nicht kann? – Zur Kartierung von Kompetenzen und Expertisen der Partner in der Task Area Editions im Bereich Consulting, Blog Post, Hypotheses, 23.02.2023. <https://textplus.hypotheses.org/3651>. [↑](#footnote-ref-17)
18. see deliverable IO4.3 [↑](#footnote-ref-18)
19. see report <https://dhd-blog.org/?p=18129> [↑](#footnote-ref-19)
20. see report <https://textplus.hypotheses.org/6074> [↑](#footnote-ref-20)
21. See  DOI: [10.5281/zenodo.12606959](https://doi.org/10.5281/zenodo.12606959) [↑](#footnote-ref-21)
22. {LINK ZUR ENTSCHEIDUNGSHILFE} [↑](#footnote-ref-22)
23. see <https://zenodo.org/doi/10.5281/zenodo.12759959> [↑](#footnote-ref-23)
24. initial post of the series: <https://textplus.hypotheses.org/10408>, post on audiovisual material: <https://textplus.hypotheses.org/10732> [↑](#footnote-ref-24)
25. See: Speer, A., Hensen, K. E., Geißler, N., Kudella, C., Sievers, M., Lemke, K., & König, S. (2024). Onboarding Guide der Task Area Editions (Version v1). Zenodo. <https://doi.org/10.5281/zenodo.10854729>. [↑](#footnote-ref-25)
26. See e.g. here: <https://textplus.hypotheses.org/10982> [↑](#footnote-ref-26)
27. Editions Tools and Software: <https://research-software.cceh.uni-koeln.de/>, Data Model and documentation: <https://gitlab.gwdg.de/textplus/curated-software-platform-for-editions>. [↑](#footnote-ref-27)
28. see appendix A in Collections’ report <https://zenodo.org/doi/10.5281/zenodo.12771255> [↑](#footnote-ref-28)
29. <https://datacite-metadata-schema.readthedocs.io/en/4.5/properties/> [↑](#footnote-ref-29)
30. All GND related blog posts of Text+: <https://textplus.hypotheses.org/?s=gnd>. [↑](#footnote-ref-30)
31. Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, Jörg Röpke: Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel 2020. text/html Format. DOI: 10.17175/2020\_006 [↑](#footnote-ref-31)
32. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html> [↑](#footnote-ref-32)
33. see Körner, et al., 2023. [↑](#footnote-ref-33)
34. See <https://text-plus.org/en/vernetzung/daten-kompetenzzentren/> and these reports: <https://zenodo.org/doi/10.5281/zenodo.12744055>; <https://zenodo.org/doi/10.5281/zenodo.12759059> [↑](#footnote-ref-34)
35. see <https://zenodo.org/doi/10.5281/zenodo.12800179> [↑](#footnote-ref-35)
36. See also: <https://text-plus.org/ueber-uns/architektur/> [↑](#footnote-ref-36)
37. See <https://text-plus.org/en/vernetzung/daten-kompetenzzentren/> for a list of all data and competence centres. [↑](#footnote-ref-37)
38. <https://text-plus.org/en/daten-dienste/dienste/> [↑](#footnote-ref-38)
39. <https://textplus.pages.gwdg.de/collections/mona-pipe/> [↑](#footnote-ref-39)
40. <https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page> [↑](#footnote-ref-40)
41. see <https://huggingface.co/LennartKeller/fiction-gbert-large-droc-np-ner>, for the complete list of components and implementations see <https://textplus.pages.gwdg.de/collections/mona-pipe/getting_started/component_overview/> [↑](#footnote-ref-41)
42. <https://registry.text-plus.org/entities/edition/>. [↑](#footnote-ref-42)
43. For a full documentation see Gradl, T., Kudella, C., Lordick, H. et al. Towards a Registry for Digital Resources – The Text+ Registry for Editions. Datenbank Spektrum 24, 151–160 (2024). <https://doi.org/10.1007/s13222-024-00479-0>. For a contextualised data model, see Measure 1 der Task Area Editions. (2024). Datenmodell Editionenregistry (Text+) (1.1). Zenodo. https://doi.org/10.5281/zenodo.12799883 [↑](#footnote-ref-43)
44. Cf. for the Classes <https://sparontologies.github.io/fabio/current/fabio.html#classes>. [↑](#footnote-ref-44)