Utilise and Preserve Text- and Language-Based Research Data

Discussion of the proposal for the funding second phase 2026–31

27.11.2025 Bonn

Funded by

Deutsche Forschungsgemeinschaft
German Research Foundation

In cooperation with

Nationale Forschungsdaten Infrastruktur

1. **What is Text+?**

2. **Where does Text+ stand now?**

3. **Text+: the road until 2031**

1. What is Text+?

2. Where does Text+ stand now?

3. Text+: the road until 2031

Covering the entire humanities research landscape
Services for the entire spectrum of text-based research
Strong combination of university and non-university institutions

Andreas Witt

Andrea Rapp

Philipp Wieder

Hanna Fischer

Philippe Genêt

21 universities, 8 academies of sciences and humanities, 12 computing centres, libraries, and non-university research institutions
6 task areas with co-spokespersons from universities and non-university research institutions:

- **Library and Archive Holdings as Research Data**: Elke Teich (Saarland University) and Philippe Genêt (DNB)
- **Language Corpora and Elicited Data**: Angelika Zirker (Tübingen University) and Andreas Witt (IDS Mannheim)
- **Lexical Resources**: Hanna Fischer (Marburg University) and Alexander Geyken (BBAW)
- **Editions**: Andrea Rapp (Technical University Darmstadt) and Andreas Speer (Academy of Sciences North-Rhine Westphalia)
- **Infrastructure Operations**: Philipp Wieder (GWDG), Vivien Petras (Humboldt University Berlin) and Regine Stein (VZG)
- **Administration**: Philipp Wieder (GWDG) and Andreas Witt (IDS Mannheim)

Covering the entire humanities research landscape
Services for the entire spectrum of text-based research
Strong combination of university and non-university institutions

Andreas Witt

Andrea Rapp

Philipp Wieder

Hanna Fischer

Philippe Genêt

21 universities, 8 academies of sciences and humanities, 12 computing centres, libraries, and non-university research institutions

6 task areas with co-spokespersons from universities and non-university research institutions:

- **Library and Archive Holdings as Research Data**: Elke Teich (Saarland University) and Philippe Genêt (DNB)
- **Language Corpora and Elicited Data**: Angelika Zirker (Tübingen University) and Andreas Witt (IDS Mannheim)
- **Lexical Resources**: Hanna Fischer (Marburg University) and Alexander Geyken (BBAW)
- **Editions**: Andrea Rapp (Technical University Darmstadt) and Andreas Speer (Academy of Sciences North-Rhine Westphalia)
- **Infrastructure Operations**: Philipp Wieder (GWDG), Vivien Petras (Humboldt University Berlin) and Regine Stein (VZG)
- **Administration**: Philipp Wieder (GWDG) and Andreas Witt (IDS Mannheim)

1. What is Text+?

2. **Where does Text+ stand now?**

3. Text+: the road until 2031

Achievements for the community (2021–26)

- Empowering community's access to text and language data
- Making research data FAIR and globally reusable
- Providing trusted services and expert consulting
- Strengthening the community through training and outreach
- Driving innovation with AI-ready data
- Leading in standardisation and authority data
- Ensuring sustainability of services
- Actively engaging in OneNFDI and Europe
- Achieving full interoperability with Base4NFDI

Achievements for the community (2021–26)

- Empowering community’s access to text and language data
- Making research data FAIR and globally reusable
- Providing trusted services and expert consulting
- Strengthening the community through training and outreach
- Driving innovation with AI-ready data
- Leading in standardisation and authority data
- Ensuring sustainability of services
- Actively engaging in OneNFDI and Europe
- Achieving full interoperability with Base4NFDI

Achievements in numbers (2021–26)

Text+ Centres
… and counting
33

Data sets via
FCS and
Registry
1.097.822

115
Tools &
services for the
community

Webportal visits
34.400

Text+ Data Space

758
Helpdesk &
consultation
requests solved

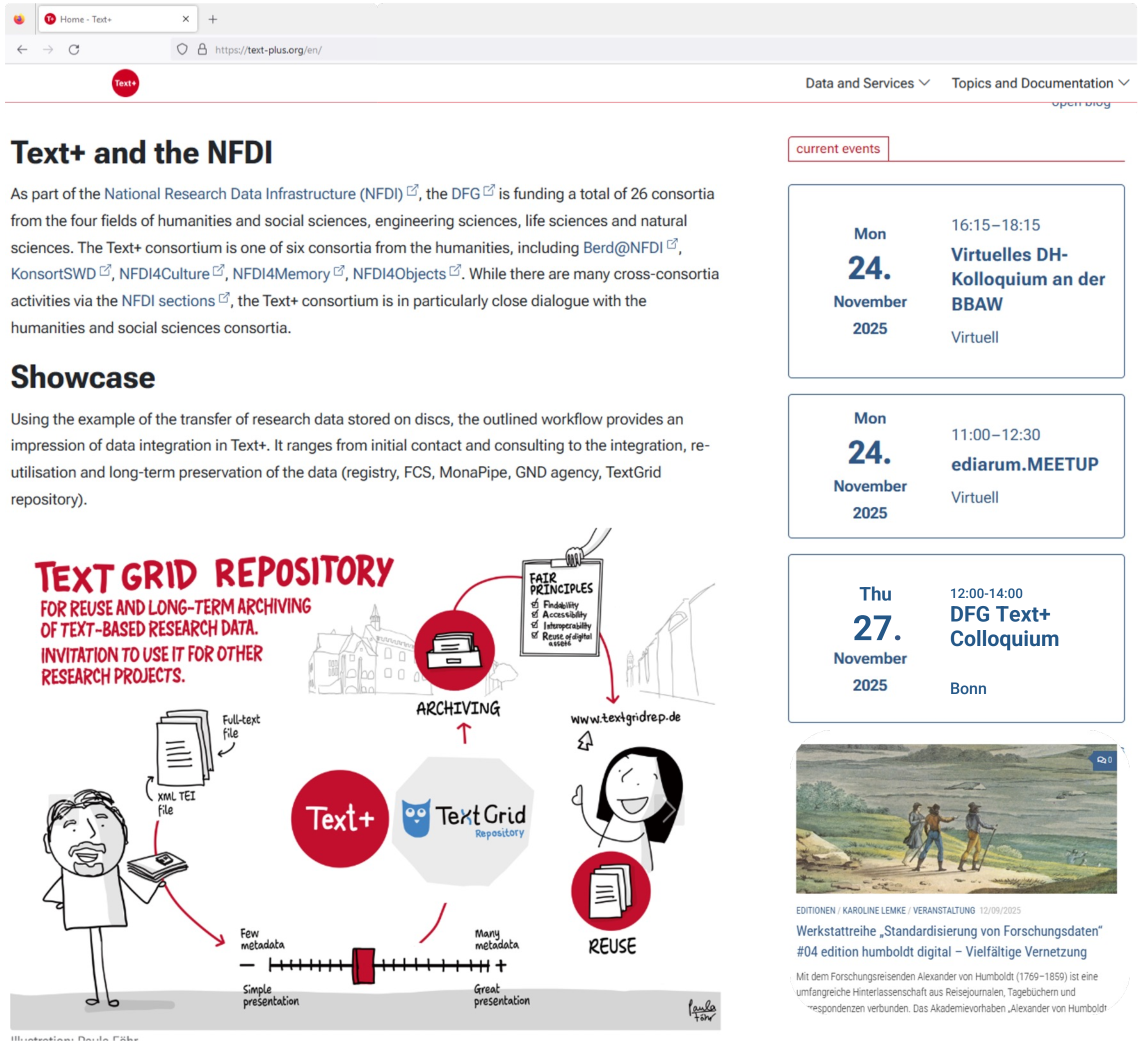
Events
… and counting
286

528
Publications via
the project
bibliography

Community activities

Text+

Text+ portal as entry point to the portfolio



Text+ and the NFDI

As part of the National Research Data Infrastructure (NFDI), the DFG is funding a total of 26 consortia from the four fields of humanities and social sciences, engineering sciences, life sciences and natural sciences. The Text+ consortium is one of six consortia from the humanities, including Berd@NFDI, KonsortSWD, NFDI4Culture, NFDI4Memory, NFDI4Objects. While there are many cross-consortia activities via the NFDI sections, the Text+ consortium is in particularly close dialogue with the humanities and social sciences consortia.

Showcase

Using the example of the transfer of research data stored on discs, the outlined workflow provides an impression of data integration in Text+. It ranges from initial contact and consulting to the integration, re-utilisation and long-term preservation of the data (registry, FCS, MonaPipe, GND agency, TextGrid repository).

current events

Mon
24.
November
2025

16:15–18:15
Virtuelles DH-Kolloquium an der BBAW
Virtuell

Mon
24.
November
2025

11:00–12:30
ediarum.MEETUP
Virtuell

Thu
27.
November
2025

12:00-14:00
DFG Text+ Colloquium
Bonn

EDITIONEN / KAROLINE LEMKE / VERANSTALTUNG 12/09/2025

Werkstattreihe „Standardisierung von Forschungsdaten" #04 edition humboldt digital – Vielfältige Vernetzung

Mit dem Forschungsreisenden Alexander von Humboldt (1769–1859) ist eine umfangreiche Hinterlassenschaft aus Reisejournalen, Tagebüchern und rrespondenzen verbunden. Das Akademievorhaben „Alexander von Humboldt

For

- Diversity of text-based academic disciplines
- Heterogenous research data encompassed and represented in the data domains

Through

- RDM services across entire research data lifecycle
- Training and consulting: fostering digital literacy
- Transfer to other disciplines and domains

Responsive to

- Tectonic shift caused by the availability of AI

Text+

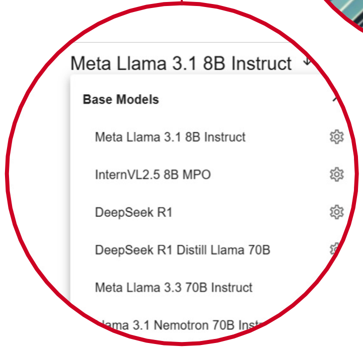
Services for the community

5 basic services integrations

base4nfdi

Data depositing form

> 20 base models in the LLM service



Meta Llama 3.1 8B Instruct

Base Models

Meta Llama 3.1 8B Instruct

InternVL2.5 8B MPO

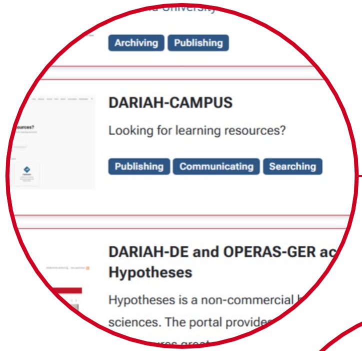
DeepSeek R1

DeepSeek R1 Distill Llama 70B

Meta Llama 3.3 70B Instruct

lama 3.1 Nemotron 70B Inst

>115 services in the catalogue



Archiving Publishing

DARIAH-CAMPUS

Looking for learning resources?

Publishing Communicating Searching

DARIAH-DE and OPERAS-GER ac Hypotheses

Hypotheses is a non-commercial sciences. The portal provide

many other services

>230 events

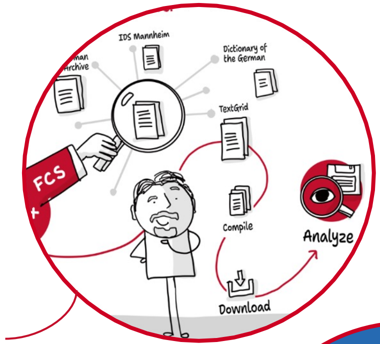
DARIAH Annual Event 2025 - The Past

Göttingen

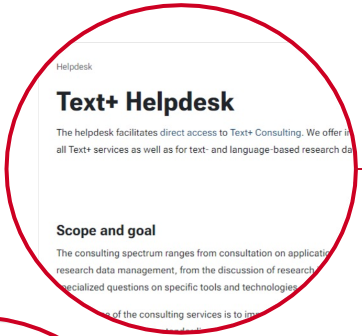
The Past

DARIAH ist eine europäische Forschungsinfrastruktur für Arts und Humanities. Ihre ist es, wissenschaftliche Gemeinschaften mit digitalen Methoden zu befähigen ultur und Gesellschaft zu schaffen, zu vernetzen und zu teilen. Die DARIAH sveranstaltungen verbinden verschiedene Formen der Begegnung DARIAH-Forschern und dem breiteren Kulturerbe, den Ku ften sowie den Computer-, Informations

FCS



Helpdesk



Helpdesk

Text+ Helpdesk

The helpdesk facilitates direct access to Text+ Consulting. We offer i all Text+ services as well as for text- and language-based research da

Scope and goal

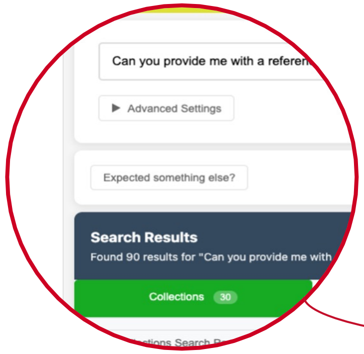
The consulting spectrum ranges from consultation on applicatio research data management, from the discussion of research pecialized questions on specific tools and technologies

e of the consulting services is to im

GND-Agency Text+

GND

AI based search (experimental)



Can you provide me with a referen

▶ Advanced Settings

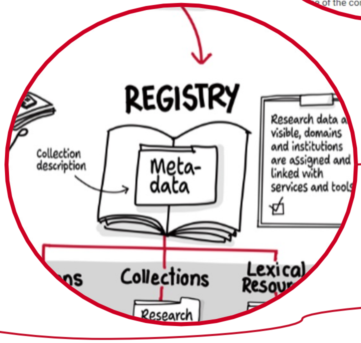
Expected something else?

Search Results

Found 90 results for "Can you provide me with

Collections 30

3 data domains in the registry


REGISTRY
Collection description
Meta-data
Research data a visible, domains and institutions are assigned and linked with services and tools
ns
Collections
Lexical Resour
Research

Text+

1. What is Text+?

2. Where does Text+ stand now?

3. **Text+: the road until 2031**

Strategic themes of Text+

- Language/text as dominant interface to data
- Dynamic interplay between structured and unstructured data
- Human-centred research culture
- Longevity of the research data infrastructure

Work programme of the data domains

M1
Federated Data Infrastructure

- Text+ Data Space
- Federated resource catalogue and Federated Content Search
- Quality Standards for Research Data (FAIR)

M2
Interfaces between Data from Data Domains

- Entity linking and authority files (GND)
- Interoperability and knowledge graph integration
- Tracing and citing data and tools

M3
Enabling Text and Language Data for AI

- Automatic annotation
- High quality and curated data for LLMs
- Derived text formats

M4
Implicit and Explicit Structures of Language Data

- Capturing structures in unstructured data
- Evaluation and proving of generative AI
- Knowledge representation of implicit and explicit structures

M5
Community Activities

- Dissemination and community outreach
- RDM training and education
- Consulting and helpdesk

Example activity:
Leveraging Library Holdings for AI

- Train models and scripts for automatic classification tasks
- Provide large scale, high quality, curated data for LLM training (copyright-free)
- Publish derived text formats (DTF) of in-copyright material

[Task Area Library and Archive Holdings as Research Data, Measure 3 Enabling text and language data for AI]

Work programme of infrastructure operations

DATA
IO-M1
Federated Research Data Infrastructure

PROCESSES
IO-M2
Service Integration w/ Data Domains

PLATFORM
IO-M3
Platform and Core Services

ECOSYSTEM
IO-M4
Coherence

- Four measures supporting two core project goals:
  - Technical gateway to the Text+ Data Space
  - Empowering text-based research
- DATA: FCS & Registry, data integration with data centres, data enrichment
- PROCESSES: UX/UI of services, technology stack coherence
- PLATFORM: enabling and supporting role for the project (e.g. LLM), Portal, Base4NFDI, IT service management
- ECOSYSTEM: alignment with NFDI, EOSC, standards, indicators and impact alignment with data domains and administration

Example activity:
Basic services integration

- IO integrates services of Base4NFDI in the architecture of Text+ such as Terminology Service (TS4NFDI), Knowledge Graph Infrastructure (KGI4NFDI), Persistent Identifiers (PID4NFDI), Jupyter4NFDI and Identity and Access Management (IAM4NFDI).
- IO aligns genuine Text+ services such as LLM or search & retrieval (FCS, registry) with the overall (upcoming) architecture of OneNFDI.
- By these activities, Text+ leverages the uptake of its services in the larger NFDI community.

[Task Area Infrastructure Operations , Measure 3 Platform and Measure 4 Ecosystem]

What Text+ will achieve until 2031

- Offering stable, reliable, sustainable services that are actively used by our communities
- Making fine-tuned LLMs and high quality, curated data for LLM training available
- Enabling automatic transcriptions and annotations
- Collaborating with specialised information services, learned societies, and other experts

[Text+ Vision] **The text- and language-oriented humanities and social sciences extensively leverage the possibilities of digitisation in their research, teaching, and transfer, establishing a common data culture.**

1. What is Text+?
2. Where does Text+ stand now?
3. Text+: the road until 2031

Thank you!

Discussion of the proposal for the second funding phase 2026–31

27.11.2025 Bonn

Andrea Rapp

Philippe Genêt

Andreas Witt

Philipp Wieder

Hanna Fischer

Funded by

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

In cooperation with

nfdi Nationale Forschungsdaten Infrastruktur

[Text+ Vision] **The text- and language-oriented humanities and social sciences extensively leverage the possibilities of digitisation in their research, teaching, and transfer, establishing a common data culture.**

1. What is Text+?

2. Where does Text+ stand now?

3. Text+: the road until 2031