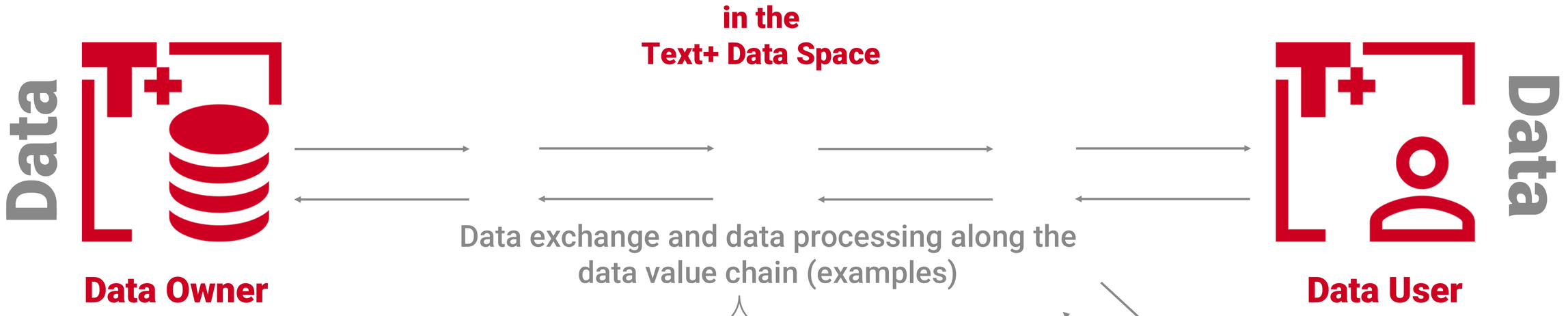
Text+ in Practice: FAIR Access to Language Corpora and Lexical Resources in a Federated Infrastructure

Collections
Lexical Resources
Editions
Infrastructure/ Operations

Thorsten Trippel – and many colleagues from the Text+ consortium
thorsten.trippel@uni-tuebingen.de

Data sharing in the Text+ Data Space



Typical Use Cases

Searching for and linking research data

Use Case: Text+ Registry

Objective: To search for distributed data using the relevant metadata

Development in Text+: Creation of the registry, integration of relevant catalogue systems, and connection to overarching structures and search systems

Status: Publicly available on the Text+ website; the database is being further expanded

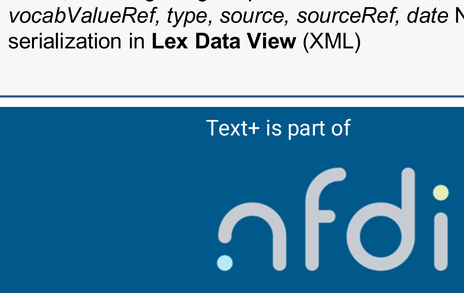
Search in research data

Use Case: Federated Content Search

Objective: Searching across distributed data, e.g. across different dictionaries

Development in Text+: Expansion of the database; extension to support queries to lexical resources

Status: Publicly available; Text+ website; database is being further expanded



Analysis of research data

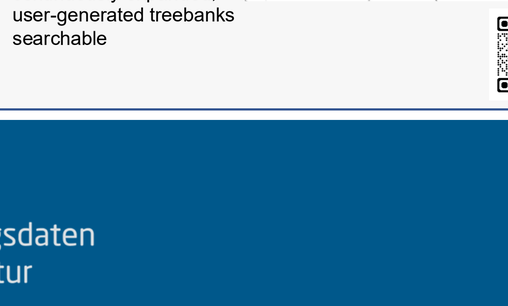
Use Case: Buchpreis@KorAP

Objective: To enable searching across all longlisted books for the German Book Prize through collaboration in Text+

Development in Text+: Automated annotation of data at the DNB, installation and ingestion of data

Status: Publicly available, linked from the German Book Prize, DNB, Text+

Further information: https://tinyurl.com/y8zrfnmk



Use of protected research data

Use Case: Derived Text Formats

Objective: To make research data subject to copyright or data protection restrictions available for reuse in research

Development in Text+: DIN 19461:2026-04 (E) standard on derived text formats, examples of use

Status: Draft national standard

Rang	N-Gramm	Häufigkeit
1	gott sei dank	43
2	ja gnädigste frau	17
3	auch heute wieder	13
4	doch auch wieder	11
5	ist doch auch	11
6	ist immer so	10
7	gnädigste frau ist	10
8	war so war	10
9	nein gnädigste frau	9
10	wird ja wohl	9
11	ist doch recht	9
12	doch immer noch	9

Häufigkeiten von 3-Grammen über mehrere Texte hinweg, bei einer Mindesthäufigkeit von 5. Beispieldaten auf der Grundlage von fünf Erzähltexten von Theodor Fontane. Schöch et al. 2020

Semantic interoperability of data

Use Case: Text+ GND Agency

Objective: To support projects in contributing their research data to the GND as standardised data.

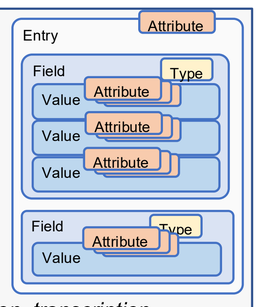
Development in Text+: Contact point: The point of contact for researchers wishing to contribute data to the GND. **Advisory service (under development):** Text+ provides advice on matters relating to the GND and makes information materials available. Data service: Support with the creation of GND-compliant standardised data. Validation, cleaning, enrichment and conversion into the GND's ingest format. Feedback to researchers regarding data enriched with a GND ID.

Status: Work in progress, but some services are already available and operational (e.g. entityXML)

GND GEMEINSAME NORMDATEI

Data Model

- Key-value based Lemma entries
  - **Entry**: a *single* result, with optional language info
  - **Field**: set of values grouped by **type**, e.g.,



entryId, lemma, phonetic, translation, transcription, definition,etymology, case, number, gender, pos, segmentation, sentiment, antonym, hyponym, hypernym, meronym, holonym, synonym, subordinate, superordinate, related, ref, senseRef, citation

- **Value**: actual "content"/value with attributes for additional context, e.g.,

xml:id, xml:lang, langUri, preferred, ref, idrefs, vocabRef, vocabValueRef, type, source, sourceRef, date Natural serialization in **Lex Data View** (XML)

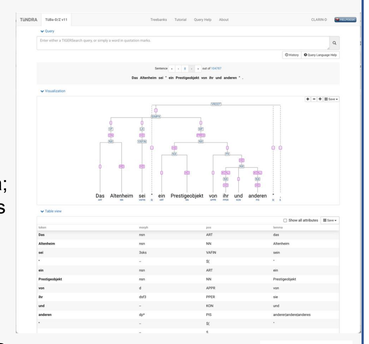
Analysis and tool integration

Use Case: Multilingual Treebank Exploration

Objective: Enable syntactic search and analysis across >400 Tübingen treebanks

Development: UD treebanks + local corpora; creation of new treebanks supported; analysis via Syndra

Status: Public tools; continuously expanded; user-generated treebanks searchable



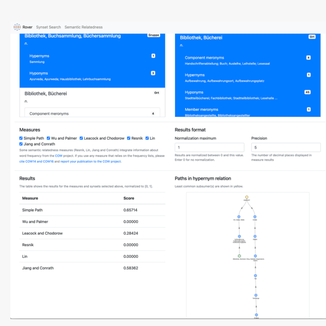
Reference data sets

Use Case: Semantic exploration of German vocabulary; Light weight ontology

Objective: Provide a structured lexical-semantic network linking >179k synsets

Development: University of Tübingen; APIs; EuroWordNet integration; GermaNet Rover

Status: Free academic access; extensive tool ecosystem; actively expanded (Release 20.0)



Funded by
DFG Deutsche Forschungsgemeinschaft German Research Foundation
The NFDI consortium Text+ is funded by Deutsche Forschungsgemeinschaft (DFG), project number 460033370.
Text+ is part of
nfdi Nationale Forschungsdaten Infrastruktur

https://www.text-plus.org