



Lexical Resources in the NFDI Project Text+

Collections
Lexical
Resources
Editions
Infrastructure/
Operations

Alexander Geyken Axel Herold

Berlin-Brandenburgische Akademie der Wissenschaften

Funded by **DFG** Deutsche Forschungsgemeinschaft
German Research Foundation

Inria-ECDF Partnership Kick-Off Workshop, June 6, 2024

Project number 460033370

Part of **nfdi** Nationale Forschungsdaten Infrastruktur



Outline

Setup of the Task Area

Specific Goals for Lexical Resources

Portfolio

Infrastructure

Where from here?

For a comprehensive general account on Text+, see Andreas Witt's talk.



Setup of the Task Area

BBAW Berlin-Brandenburg Academy of Sciences and Humanities (lead)

IDS Leibniz Institute for the German Language

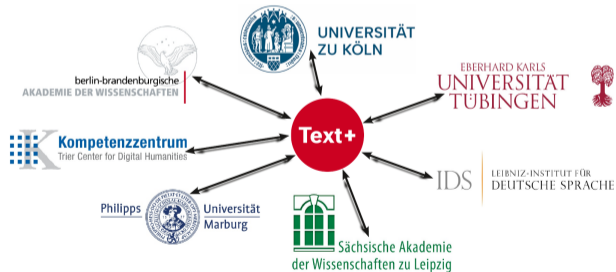
SAW Saxon Academy of Sciences and Humanities

DSA Research Center Deutscher Sprachatlas

UniTr University of Trier, Trier Center for Digital Humanities

UniTü University of Tübingen

UniK University of Cologne



Specific Goals for Lexical Resources

- ▶ provide FAIR resources
 - ▶ metadata curation
 - ▶ lexical data curation
 - ▶ persistent data storage
- ▶ create a common environment to explore and use lexical resources
- ▶ integrate with broader NFDI infrastructure
- ▶ be useful across the Humanities for a broad range of research
- ▶ provide consulting for scientists (and members of the public)



Portfolio

Three thematic clusters

- ▶ German dictionaries in a European context
- ▶ born-digital lexical resources
- ▶ non-Latin scripts



Portfolio

Three thematic clusters

- ▶ German dictionaries in a European context
- ▶ born-digital lexical resources
- ▶ non-Latin scripts

Three main data sources

- ▶ consortial partners' resources
- ▶ short-term funded partners' resources (annual calls)
- ▶ external partners' resources (in focus for second phase)



Consortial partners' resources

- ▶ dictionaries of present day German (BBAW, IDS, UniTü, SAW)
- ▶ dictionaries of other language stages of German (BBAW, UniTr, SAW)
- ▶ historic dictionaries (UniTr, BBAW)
- ▶ dialect dictionaries (DSA, UniTr)
- ▶ dictionaries of ancient languages (UniK)
- ▶ specialized smaller scale dictionaries (IDS)
- ▶ statistical data (BBAW, SAW)

The image shows a screenshot of the GermanNet interface. At the top, it displays 'Typische Verbindungen zu Wort: [berechnet]' with a list of words: Bedauern, Geste, schwir, Sinn, Tat, böse. Below this, there are search filters for 'deutsch', 'englisch', 'spanisch', 'finden', 'deutsch', 'gefügelt', 'gesprochen', 'klar', 'lateinisch', 'letzt', 'lobend', 'einzig', 'paar', 'sagen', 'scharf', 'sprechen', 'was', 'wann'. On the right, there is a network diagram labeled 'GermanNet' and a list of resources with their respective categories and counts. A legend on the left lists various linguistic resources with their corresponding colors: Peronymwörterbuch (red), Spracherbuch (orange), Feste Wortverbindungen (yellow), Kommunikationsverben (green), Vorlauf Formen (purple), Fremdwörterbuch (blue), Neologismenwörterbuch (teal), Demokratiediskurs 1918-25 (dark green), and Schuldiskurs 1945-55 (light green). At the bottom left, there are logos for 'RDF' and 'TEI'. The main content area shows a search result for 'Wort' with a count of 1,800 and a list of related words and phrases. A 'Bedeutung:' section provides a detailed explanation of the word's usage and historical context.



Portfolio

Short term funded partners' resources

- ▶ Middle High German Dictionary (Trier/Göttingen, 2023)
- ▶ Lower Sorbian dictionaries (Bautzen/Cottbus, 2023)
- ▶ Corpus Glossariorum Latinorum (München, 2023)
- ▶ Thesaurus Linguae Aegyptiae (Berlin, ongoing)

—→ in support of the three thematic clusters



Infrastructure

Main components and areas of work

Registry metadata on all things Text+ (resources, services, repositories, ...)

Repositories provided by all Text+ partners
thematically specialized
open for depositing FAIR data

LexFCS *federated content search*
query lexical resources across different types of data representation

LLOD linguistic linked open data
(experimental stage)

Consulting keeping in touch with the community
provide help and support



Infrastructure

Lexical Federated Content Search (LexFCS)

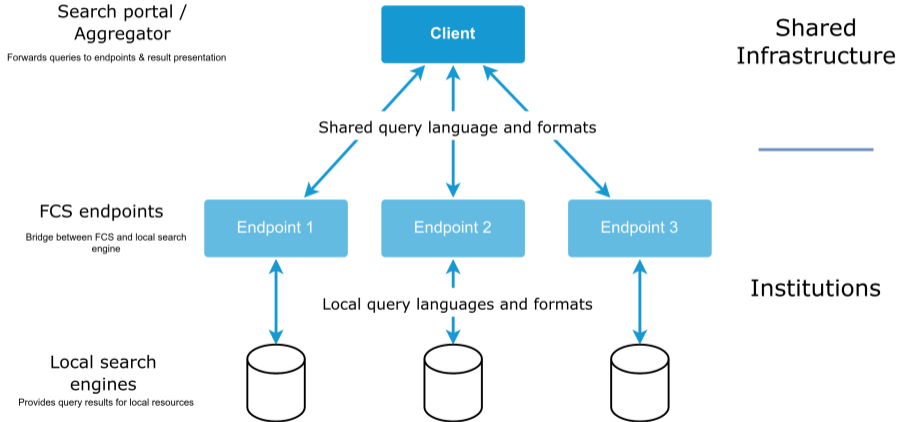
- ▶ based on CLARIN FCS (for annotated text corpora)
- ▶ different underlying primary data model:
sequence of annotated tokens (FCS)
→ tree structures or general graphs (LexFCS, a specialized *data view*)
- ▶ several (common) serialization formats for lexical data:
TEI, TEI Lex-0, RDF, custom formats
 - ▶ challenge: generalize all data representations for common querying
 - ▶ challenge: transform return data into common format for aggregation

Eckart/Herold/Körner/Wiegand (2023): [A federated search and retrieval platform for lexical resources in Text+ and CLARIN](#). In: Medved' et al. (eds.): *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, pp. 280–292



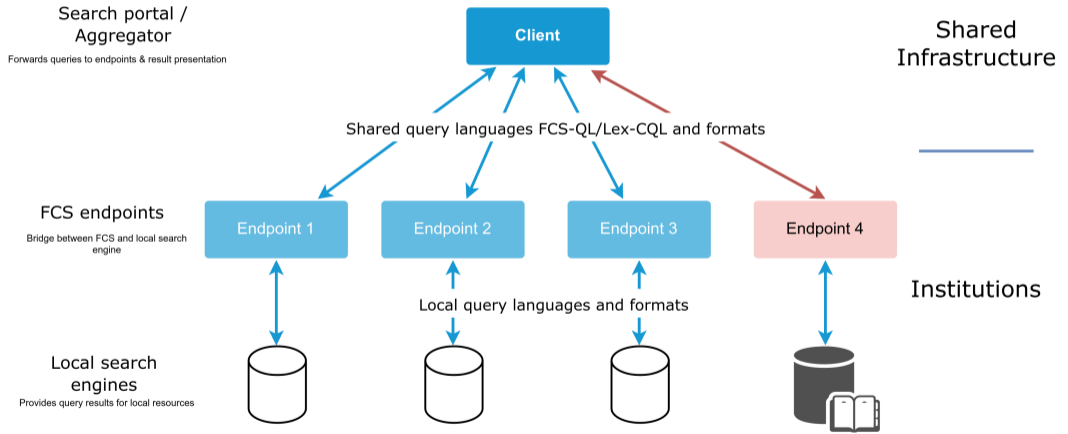
Infrastructure

FCS Architecture



Infrastructure

FCS Architecture with LexFCS extension



Infrastructure

Underlying data, EtymWB, TEI

```
<entry type="main" xml:id="E_q_9">
  <form type="headword">
    <orth extent="full">Qual</orth>
    <gramGrp>
      <pos value="N"></pos>
      <gen value="feminine">f.</gen>
    </gramGrp>
  </form>
  <div type="etym" xml:id="W_Q_9">
    <!-- ... -->
  </div>
</entry>
```



Infrastructure

Underlying data, GermaNet 11, idiosyncratic

```
<synset id="s26576" category="nomen" class="Koerper">  
  <lexUnit id="l36322" sense="1" source="core"  
    namedEntity="no" artificial="no" styleMarking="no">  
    <orthForm>Qual</orthForm>  
  </lexUnit>  
  <!-- ... -->  
</synset>
```

```
<wiktionaryParaphrase lexUnitId="l36322"  
  wiktionaryId="w43520" wiktionarySenseId="0"  
  wiktionarySense="Schmerz, Leid, etwas physisch oder  
  psychisch Belastendes" edited="no" />
```



Infrastructure

From data to query

- ▶ identify common low level lexical “data types”:
 - (unspecified) full text search
 - lemma headword of an entry
 - pos part-of-speech, word class
 - def definition, semantic paraphrase
 - xr resource internal semantic relation (work in progress)
 - senseRef links to external authority files (draft)
- ▶ LexCQL specification (fields, operators, return values)
- ▶ local search engines translate LexSQL into local query language
- ▶ local search engines transform results into TEI Lex-0

→ <https://fcs.text-plus.org/>



Infrastructure

LexFCS Aggregator

Lex CQL query

Search for

Any Language ▾

Lexical Contextual Query Language (Lex-CQL) ▾

in

31 selected resources ▾

and show up to

10

hits per endpoint



LexFCS Aggregator

25 matching resources found

31 resources searched




Display as Key Word In Context

 Download ▾



Mittelhochdeutsches Wörterbuch – Akademie der Wissenschaften und der Literatur Mainz und Akademie der Wissenschaften zu Göttingen

 View











-  eiserne Rüstung (Ringpanzerdecke?) eines **Elefanten**:
-  umschlossene und überdachte Anlage (vgl. Rosenqvist 1,128, Suolahti S.119f.) auf **Elefanten** (in Bildern des 13.-15. Jhs. sowohl als mehrere Krieger schützender Verschlag wie als ...
-  'Elfenbein', der aus den Stoßzähnen des **Elefanten** gewonnene kostbare Werkstoff sakraler und profaner Elfenbeinschnitzerei:



LexFCS Aggregator

▼ Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm – Uni Trier - Wörterbuchnetz

 View











-  konnten. Winkelmann 3, 51 ; zu seiner bedeckung folgt auf einem **elefanten** mit eisenblechen behängt der riese Moulineau. Wieland 4, 14 ; er
-  eine achse bewegen. man dreht den bratenwender, das rad. der **elefant** kann die (walze der) orgel mit dem rüssel drehen. trien
-  **elefant** , m. elephas, ahd. **elefant** und helfant, ags. ylpent und verkürzt
-  **elefantenaue** , n. nennt man ein bei starker geschwulst hervortretendes auge.
-  **elefantenfell** , n. des negers schild ist eine spanne dick und ganz
-  **elefantengerippe** , n. in der that kleidet er die **elefantengerippe** der götterlehre
-  **elefantenkalt** , n. natus ex elephanto: tretet nicht so mastig auf, wie
-  **elefantenknochen** , m.
-  **elefantenaus** , f. anacardium, ein indischer baum mit nierenartiger nusz.
-  **elefantensark** , n. hätte sie nicht als fürstin verlangen können z. b.,



LexFCS Aggregator

▼ Digital Dictionary of the German Language – Berlin-Brandenburg Academy of Sciences and Humanities (Lexical Resources)

 View

-  Lemma Elefantenfriedhof : POS NOUN . Def. oft von Menschen geheiligter Ort, wo angeblich Elefanten zum Sterben hingehen
-  Lemma aus einer Mücke einen Elefanten machen : POS X . [umgangssprachlich] Def. (jmd. macht aus einer Mücke einen Elefanten) etw. hochspielen, (künstlich) aufbauschen, dramatisieren; aus einer kleinen, unbedeutenden Angelegenheit eine (übertrieben) große machen
-  Lemma Elefantenbaby : POS NOUN . Def. Kalb bzw. Junges des Elefanten kurz nach der Geburt
-  Lemma Babyelefant : POS NOUN . Def. sehr junger Elefant
-  Lemma Elefant im Porzellanladen : POS X . Def. (wie ein, der Elefant im Porzellanladen (= unbeholfen, grob, rücksichtslos; sehr ungeschickt, tollpatschig))
-  Lemma Mastodon : POS NOUN . Def. ausgestorbenes Rüsseltier des Tertiärs; Das Mastodon ist Vorfahre des heutigen Elefanten.
-  Lemma Elefantenrennen : POS NOUN . Def. (besonders in Südasien veranstaltetes) Wettrennen zwischen dressierten Elefanten
-  Lemma Leitbulle : POS NOUN . Def. ranghöchstes männliches Tier in einer Herde Rinder, Elefanten, Hirsche, Wale o. Ä.
-  Lemma Elfenbeinschnitzer : POS NOUN . Def. Kunsthandwerker, der (berufsmäßig) Schnitzereien aus (Mammut-)Elfenbein, Horn o. ä. Materialien fertigt; Nach dem Washingtoner Artenschutzübereinkommen von 1973 ist der Handel mit Elfenbein (von Elefanten und Narwalen) verboten oder stark eingeschränkt, so dass in der EU hauptsächlich fossiles Elfenbein (meist Mammutelfenbein) für die Elfenbeinschnitzerei verwendet wird.
-  Lemma Elefantengruppe : POS NOUN . Def. Menge zusammenlebender oder gemeinsam (im Zirkus) auftretender Elefanten



Infrastructure

Observations, summary

- ▶ (somewhat) flexible interpretation of operators (e. g. wrt capitalization)
- ▶ should be more rigid wrt return data types
(complete article, specific part of an article?)
- ▶ provide complete mark-up or just highlight the matches?

more fundamentally:

- ▶ trade-off between *generalization* of diverse dictionary data and highly *targeted querying*
- ▶ so far, mostly German resources (currently 31)



Where from here?

Future work on LexFCS/LexCQL

- ▶ integrate more lexical resources
- ▶ map more basic lexical data types
(e. g. time, space, usage labels, citations, translation equivalents, ...)
- ▶ adapt for non-latin scripts
- ▶ adapt for external authority files
(interface to other NFDI consortia)
- ▶ adapt for hierarchical dependencies?
- ▶ nicer ways to present results than just a flat list?





Thank you!

Collections
Lexical
Resources
Editions
Infrastructure/
Operations

This presentation was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e. V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

Funded by

DFG Deutsche
Forschungsgemeinschaft
German Research Foundation

Project number 460033370

Part of

nfdi Nationale
Forschungsdaten
Infrastruktur

<https://www.text-plus.org>

