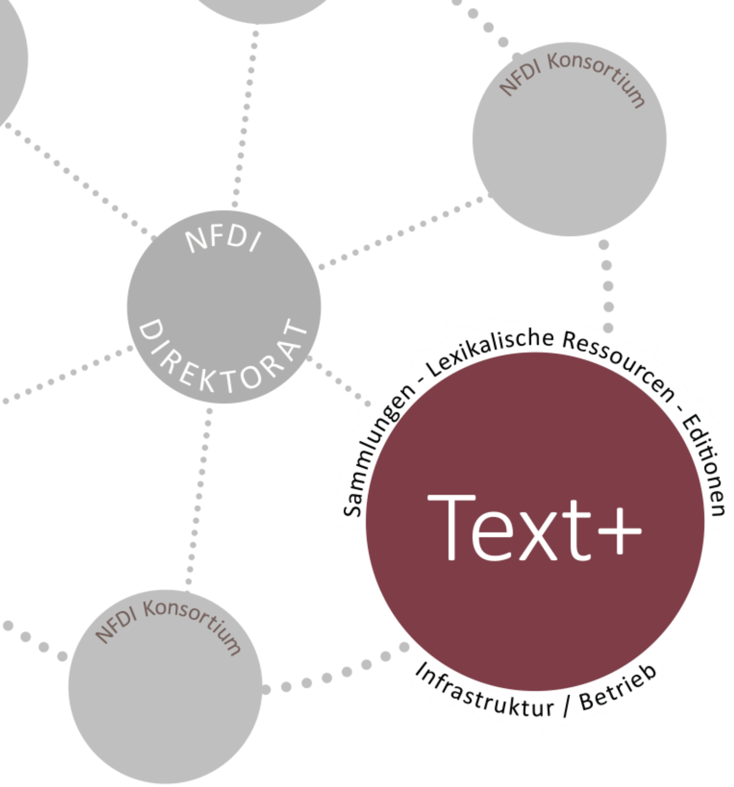
Presented by SAW and BBAW in the context of Text+. The NFDI consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG) - project number 460033370.



eLex 2023

A Federated Search and Retrieval Platform for Lexical Resources in Text+ and CLARIN

Thomas Eckart, Axel Herold, **Erik Körner**, Frank Wiegand
Saxon Academy of Sciences and Humanities in Leipzig, Leipzig, Germany
Berlin-Brandenburg Academy of Sciences and Humanities, Berlin, Germany
{eckart,koerner}@saw-leipzig.de, {herold,wiegand}@bbaw.de

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Sächsische Akademie der Wissenschaften zu Leipzig

Outline

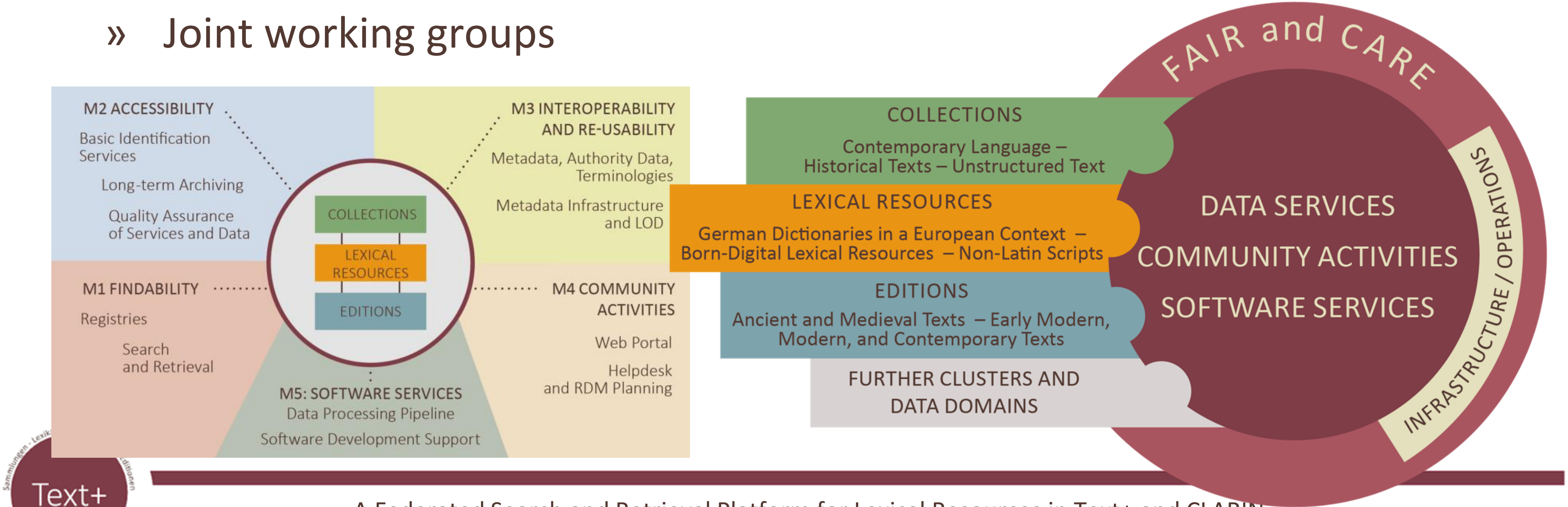
- Text+ and Lexical Resources
- Federated Content Search Infrastructure
- FCS Specification Extension for Lexical Resources
- Next Steps and Future Work

Text+ and NFDI

- Text+: research data consortium focused on **language and text data**
- part of Germany's National Research Data Infrastructure (NFDI)
  - Aims: make research data **available** for scientific usage, support their **interlinkage**, and their long-term **preservation**
  - inter-disciplinary network of data and services based on common standards and the FAIR principles between consortia from various research areas

Text+ Data Domains

- 3 data domains: Collections, Lexical Resources, Editions
- Joint working groups



M2 ACCESSIBILITY

Basic Identification Services

Long-term Archiving

Quality Assurance of Services and Data

M3 INTEROPERABILITY AND RE-USABILITY

Metadata, Authority Data, Terminologies

Metadata Infrastructure and LOD

COLLECTIONS

LEXICAL RESOURCES

EDITIONS

M1 FINDABILITY

Registries

Search and Retrieval

M4 COMMUNITY ACTIVITIES

Web Portal

Helpdesk and RDM Planning

M5: SOFTWARE SERVICES

Data Processing Pipeline

Software Development Support

FAIR and CARE

COLLECTIONS

Contemporary Language – Historical Texts – Unstructured Text

LEXICAL RESOURCES

German Dictionaries in a European Context – Born-Digital Lexical Resources – Non-Latin Scripts

EDITIONS

Ancient and Medieval Texts – Early Modern, Modern, and Contemporary Texts

FURTHER CLUSTERS AND DATA DOMAINS

DATA SERVICES

COMMUNITY ACTIVITIES

SOFTWARE SERVICES

INFRASTRUCTURE / OPERATIONS

Data Domain Lexical Resources

- Thematic clusters
  - German dictionaries in the European context
  - Born-digital lexical resources
  - Non-latin scripts and under-resourced languages

- Resource types
  - Dictionaries (mono/bilinigual)
  - Encyclopedias
  - Normative Data (GND, …)
  - Terminology Databases
  - Ontologies, Knowledge Databases
  - Word Nets (Princeton, GermaNet)
  - …

Data Domain Lexical Resources
Data Formats

- Wide, diverse spread of formats with custom search functionalities
  - Generic and customized TEI/XML to legacy XML formats
  - Table-like serializations (lemma lists, frequency information)
  - Custom, proprietary formats
  - Geographic information (images of maps), Character Sets
  - …
- Challenges due to heterogeneity for unified representation for search and retrieval

Data Domain Lexical Resources Findability

- **Decentralized** dictionary platform, **federated** approach
  - Heterogeneous nature of resources, formats, annotation levels, technical architectures
- CLARIN *Federated Content Search (FCS)*
  - Framework for accessing spatially distributed text corpora
  - Common specification of techn. interfaces, data formats, query languages

Other Approaches for Linking Lexical Resources

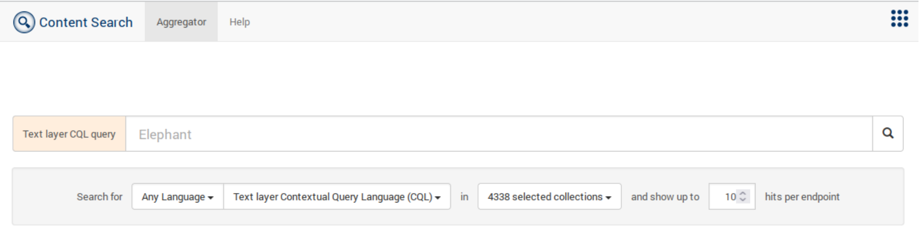
- Often organisationally restricted, e.g. “Wörterbuchnetz” by Trier, Global WordNet Association
- Initiatives in common research infrastructures (CLARIN, DARIAH); ELEXIS
- Standardized formats, e.g. TEI, refinements by DARIAH, ELEXIS; RDF/OntoLex, …
- Collaborative / not exclusively academic dictionaries, e.g. Wiktionary, DBPedia, Wikidata

Why Federated Content Search?

- Advantages to endpoints (data centers)
  - Full control about access to resources
    - Copyright, licensing or data protection
  - Knowledge about resource, e.g. how to search, rank results, ...
  - Visibility
- Advantages to end users
  - Ease of use and simple overview of resources and results
  - Simple search box – „Just like google“
  - For details and expert search options → backlink to endpoint

CLARIN Federated Content Search (FCS)

- Federated “Corpus Query Platform”
- FCS =
  - RESTful protocol
  - Query languages & data formats
  - Data Aggregator + web portal
  - (Software ecosystem)

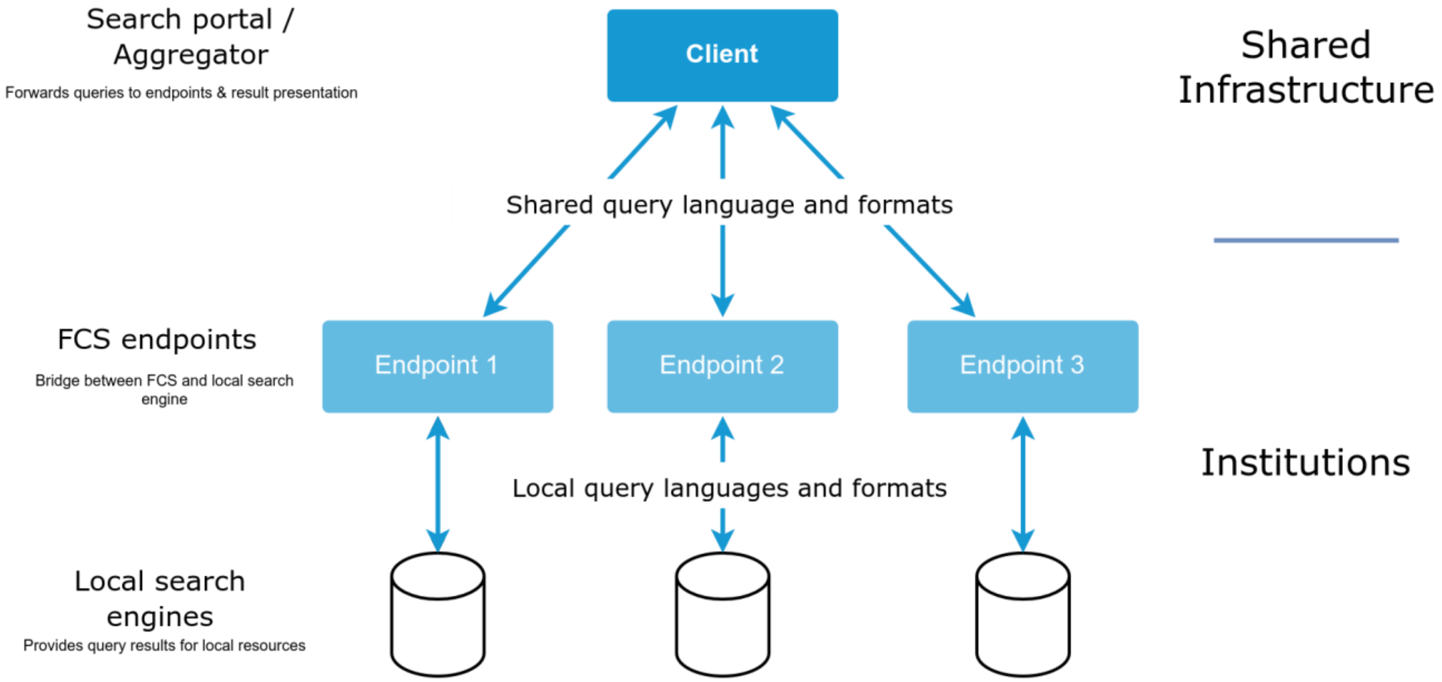


Text layer CQL query: Elephant

Search for Any Language Text layer Contextual Query Language (CQL) in 4338 selected collections and show up to 10 hits per endpoint

- FCS ≠ complete replacement for local search engines

CLARIN Federated Content Search (FCS)


Search portal /
Aggregator
Forwards queries to endpoints & result presentation
Client
Shared Infrastructure
Shared query language and formats
FCS endpoints
Bridge between FCS and local search engine
Endpoint 1
Endpoint 2
Endpoint 3
Local query languages and formats
Institutions
Local search engines
Provides query results for local resources

Communication Protocol: FCS Core 2.0

- Extension of SRU (Search/Retrieval via URL) / searchRetrieve
  - Standardized by Library of Congress LoC / OASIS
  - Data as XML
- RESTful
  - **Explain**: Existing resources
    - Language, annotations, supported data formats, etc.
  - **SearchRetrieve**: search query
- **FCS-QL** as general query language (FCS 2.0)

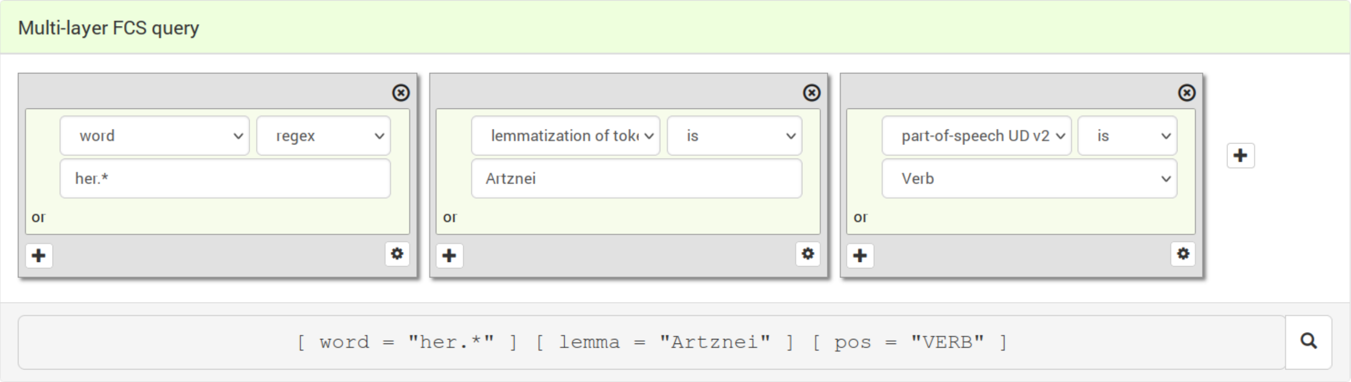
Assumption over data structure

» Full text + optional annotation layers

Full Text	*Die (The)*	*Autos (cars)*	*Sind (are)*	*Schnell (fast)*
Part of Speech (UD17)	*DET*	*NOUN*	*VERB*	*ADJ*
Base form	*Das (The)*	*Auto (car)*	*Ist (is)*	*Schnell (fast)*
Phonetic transcription (=SAMPA)	…	…	…	…
Orthographic transcription	…	…	…	…
Orthographic normalization	…	…	…	…

Query Language FCS-QL

- Similar to CQP (e.g. corpus query workbench)
- Supports multiple annotation layers



Multi-layer FCS query

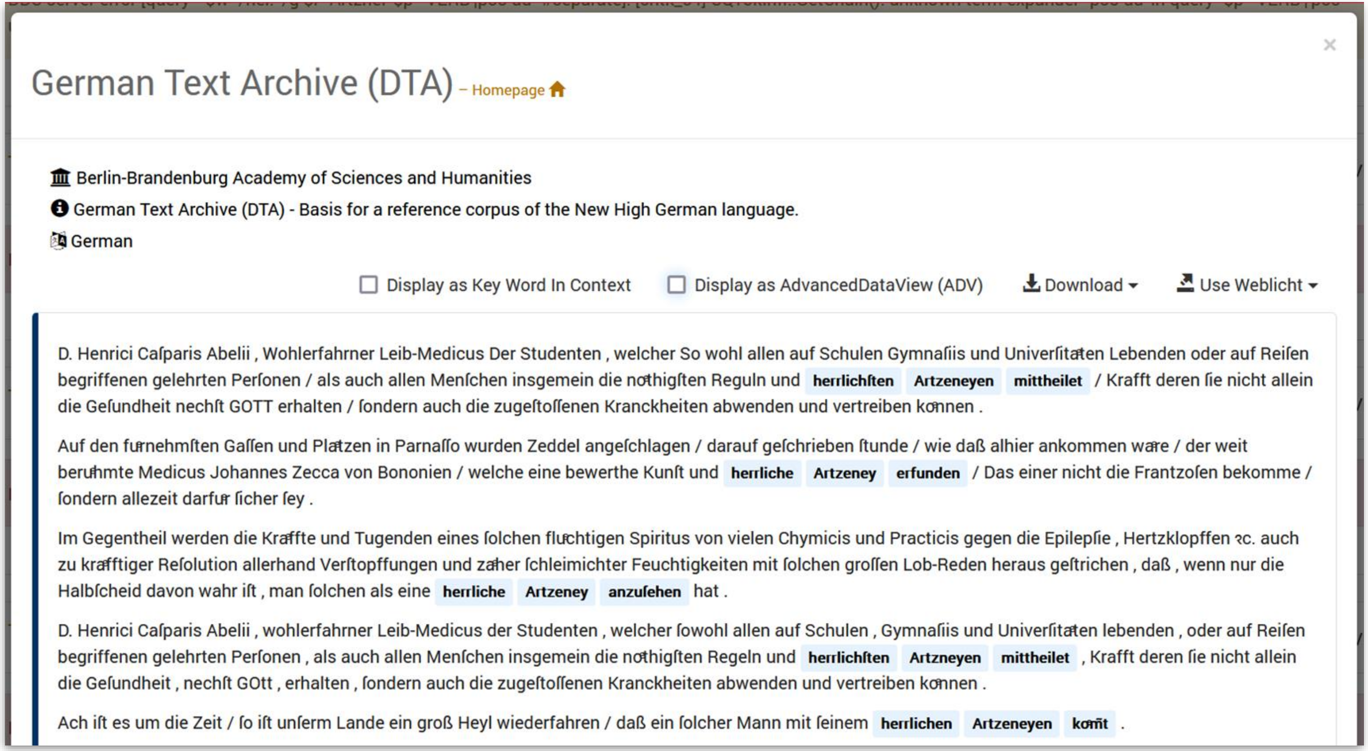
word regex her.*

lemmatization of tok is Artznei

part-of-speech UD v2 is Verb

```
[ word = "her.*" ] [ lemma = "Artznei" ] [ pos = "VERB" ]
```

Visualization of results (1/2)



German Text Archive (DTA) – Homepage

Berlin-Brandenburg Academy of Sciences and Humanities

German Text Archive (DTA) - Basis for a reference corpus of the New High German language.

German

☐ Display as Key Word In Context
☐ Display as AdvancedDataView (ADV)
Download ▾
Use Weblicht ▾
×

D. Henrici Caſparis Abelii , Wohlerfahrner Leib-Medicus Der Studenten , welcher So wohl allen auf Schulen Gymnaſiis und Univerſitaͤten Lebenden oder auf Reiſen begriffenen gelehrten Perſonen / als auch allen Menſchen insgemein die noͤthigſten Reguln und **herrlichſten** **Artzeneyen** **mittheilet** / Krafft deren ſie nicht allein die Geſundheit nechſt GOTT erhalten / ſondern auch die zugeſtoſſenen Kranckheiten abwenden und vertreiben koͤnnen .

Auf den fuͤrnehmſten Gaſſen und Plaͤtzen in Parnaſſo wurden Zeddel angeſchlagen / darauf geſchrieben ſtunde / wie daß alhier ankommen waͤre / der weit beruͤhmte Medicus Johannes Zecca von Bononien / welche eine bewerthe Kunſt und **herrliche** **Artzeney** **erfunden** / Das einer nicht die Frantzoſen bekomme / ſondern allezeit darfuͤr ſicher ſey .

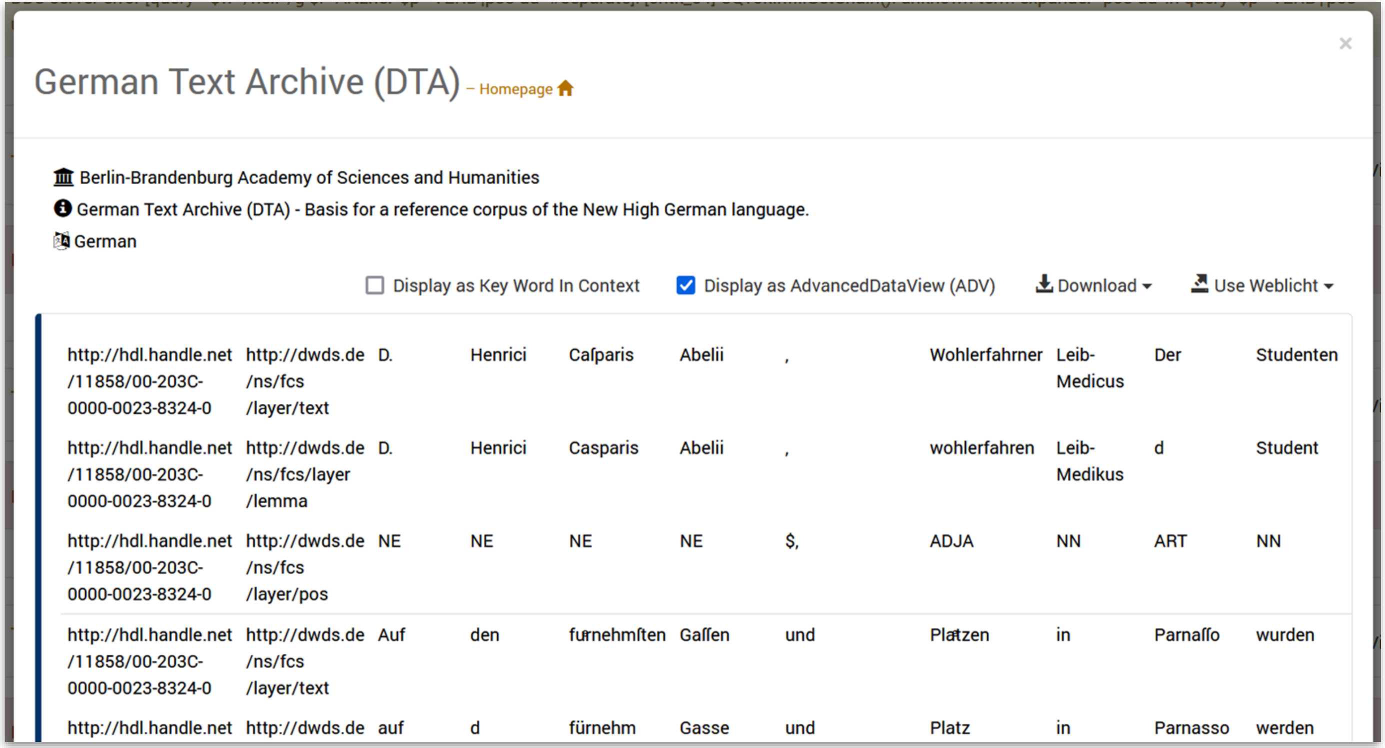
Im Gegentheil werden die Kraͤffte und Tugenden eines ſolchen fluͤchtigen Spiritus von vielen Chymicis und Practicis gegen die Epilepſie , Hertzklopffen ꝛc. auch zu kraͤfftiger Reſolution allerhand Verſtopffungen und zaͤher ſchleimichter Feuchtigkeiten mit ſolchen groſſen Lob-Reden heraus geſtrichen , daß , wenn nur die Halbſcheid davon wahr iſt , man ſolchen als eine **herrliche** **Artzeney** **anzuſehen** hat .

D. Henrici Caſparis Abelii , wohlerfahrner Leib-Medicus der Studenten , welcher ſowohl allen auf Schulen , Gymnaſiis und Univerſitaͤten lebenden , oder auf Reiſen begriffenen gelehrten Perſonen , als auch allen Menſchen insgemein die noͤthigſten Regeln und **herrlichſten** **Artzneyen** **mittheilet** , Krafft deren ſie nicht allein die Geſundheit , nechſt GOtt , erhalten , ſondern auch die zugeſtoſſenen Kranckheiten abwenden und vertreiben koͤnnen .

Ach iſt es um die Zeit / ſo iſt unſerm Lande ein groß Heyl wiederfahren / daß ein ſolcher Mann mit ſeinem **herrlichen** **Artzeneyen** **kom̃t** .

API-Call

Visualization of results (2/2)



German Text Archive (DTA) – Homepage

Berlin-Brandenburg Academy of Sciences and Humanities

German Text Archive (DTA) - Basis for a reference corpus of the New High German language.

German

□ Display as Key Word In Context ☑ Display as AdvancedDataView (ADV) Download ▾ Use Weblicht ▾

http://hdl.handle.net/11858/00-203C-0000-0023-8324-0	http://dwds.de/ns/fcs/layer/text	D.	Henrici	Caſparis	Abelii	,	Wohlerfahrner	Leib-Medicus	Der	Studenten
http://hdl.handle.net/11858/00-203C-0000-0023-8324-0	http://dwds.de/ns/fcs/layer/lemma	D.	Henrici	Casparis	Abelii	,	wohlerfahren	Leib-Medikus	d	Student
http://hdl.handle.net/11858/00-203C-0000-0023-8324-0	http://dwds.de/ns/fcs/layer/pos	NE	NE	NE	NE	$,	ADJA	NN	ART	NN
http://hdl.handle.net/11858/00-203C-0000-0023-8324-0	http://dwds.de/ns/fcs/layer/text	Auf	den	fuͤrnehmſten	Gaſſen	und	Plaͤtzen	in	Parnaſſo	wurden
http://hdl.handle.net	http://dwds.de	auf	d	fürnehm	Gasse	und	Platz	in	Parnasso	werden

Existing Resources

- Ecosystem
  - Reference implementations
  - Libraries (Java), Documentation
  - FCS/SRU Validator
  - Endpoint Registry
- Existing resources (actual number fluctuate)
  - 20 institutes in 38 endpoints in 11 countries
  - About 200 „collections“ in varying granularity in ~60 languages

FCS for Lexical Resources?

» No – not yet!

» Focus on text „streams“ (corpora, transcriptions) & querying of annotation layers

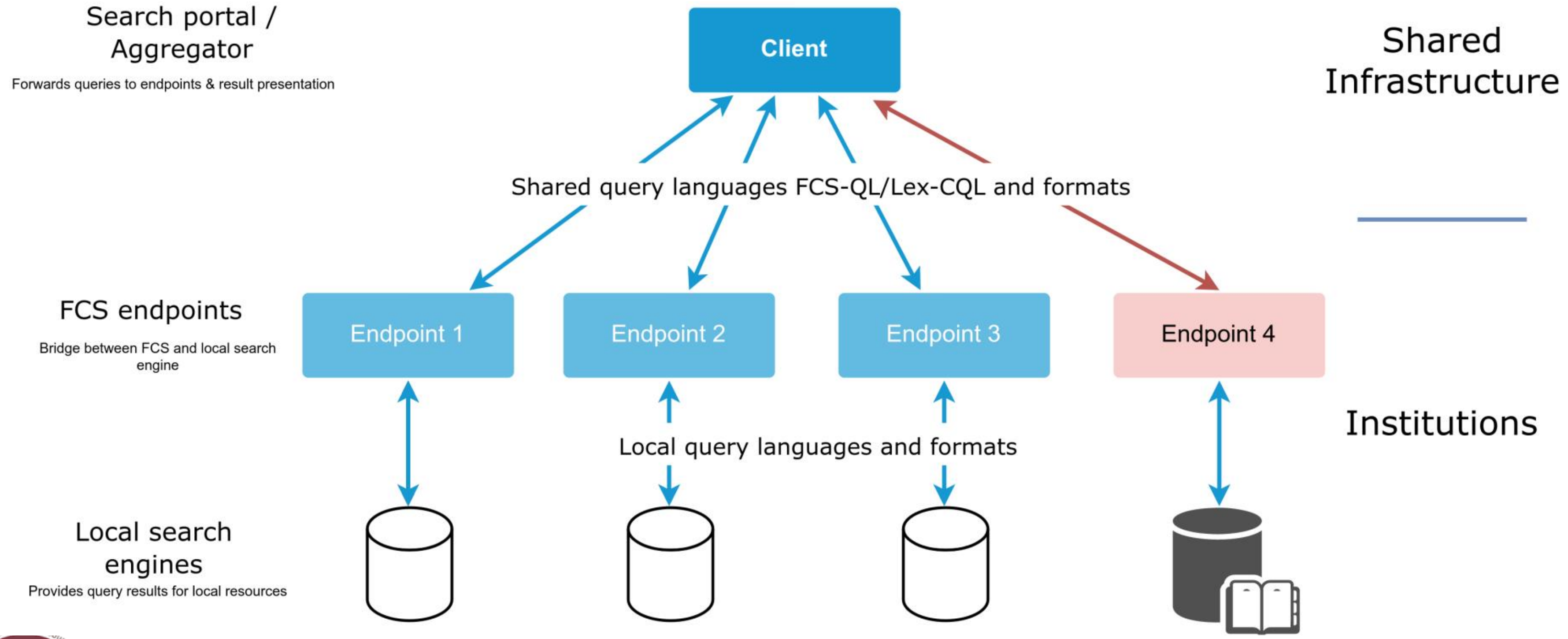
» lexical resources structurally completely different (word lists, word nets/graphs, key-value based, …)

⇨ Specification extension

FCS Specification Extension for Lexical Resources Goals

- Query language dedicated to querying lexical entries
  - Subset of **Contextual Query Language** (CQL), agreements on accessible fields of information for a lexeme, complex queries
- Common data formats for unified result presentation
  - Mandatory **LexHITS** data view with **inline annotation** of information types
  - Advanced tabular representation, key-value style
- Compatibility with FCS architecture
  - Reuse of features: access control for restricted resources, automatic registering of endpoints

FCS Specification Extension for Lexical Resources


Search portal / Aggregator
Forwards queries to endpoints & result presentation
Client
Shared Infrastructure
Shared query languages FCS-QL/Lex-CQL and formats
FCS endpoints
Bridge between FCS and local search engine
Endpoint 1
Endpoint 2
Endpoint 3
Endpoint 4
Institutions
Local query languages and formats
Local search engines
Provides query results for local resources

FCS Specification Extension for Lexical Resources Draft v0.1

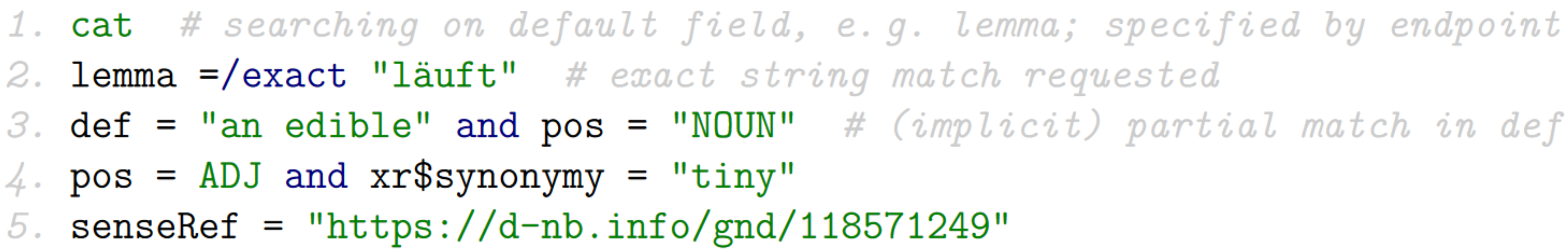
- Query Language – LexCQL
- DataViews – LexHITS + Tabular (draft)

- Draft: https://doi.org/10.5281/zenodo.7849753

LexCQL

- Subset of Contextual Query Language (CQL)
- Relation „=“
- Operators AND/OR/NOT
- (draft) Fuzzy with „/exact“ Modifier

- Fields:
  - lemma,
  - pos (UD17)
  - def
  - xr$synonymy, xr$hyponymy, …
  - (draft) senseRef



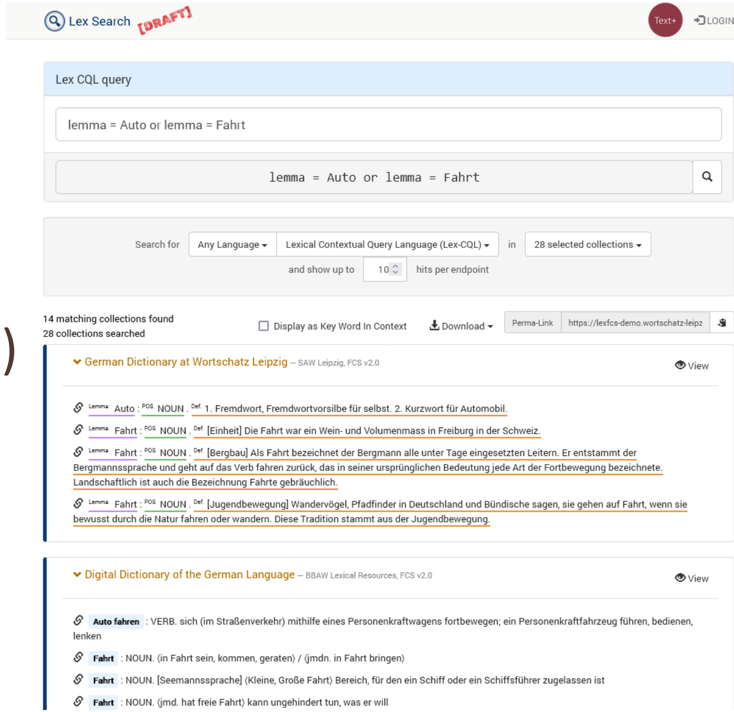
```
cat  # searching on default field, e.g. lemma; specified by endpoint
lemma =/exact "läuft"  # exact string match requested
def = "an edible" and pos = "NOUN"  # (implicit) partial match in def
pos = ADJ and xr$synonymy = "tiny"
senseRef = "https://d-nb.info/gnd/118571249"
```

LexHITS

- » Seamless integration for *mandatory* result format (HITS) BUT optional annotation of „lemma“, „pos“ and „def“
  → visual hints in frontend (e.g. Aggregator)

```
<fcs:DataView type="application/x-textplus-fcs-hits+xml">
  <hits:Result xmlns:hits="http://textplus.org/fcs/dataview/hits"
    <hits:Hit kind="lex-lemma">Apple</hits:Hit>:
    <hits:Hit kind="lex-pos">NOUN</hits:Hit>.
    <hits:Hit kind="lex-def">An apple is an edible fruit produced by
an apple tree.</hits:Hit>
  </hits:Result>
</fcs:DataView>
```

Apple: NOUN. An apple is an edible fruit produced by an apple tree.



Lex CQL query

lemma = Auto or lemma = Fahrt

`lemma = Auto or lemma = Fahrt`

Search for Any Language ▾ Lexical Contextual Query Language (Lex-CQL) ▾ in 28 selected collections ▾
and show up to 10 hits per endpoint

14 matching collections found
28 collections searched

☐ Display as Key Word In Context Download ▾ Perma-Link https://lexfcs-demo.wortschatz-leipz

German Dictionary at Wortschatz Leipzig – SAW Leipzig, FCS v2.0 View

Lemma Auto : POS NOUN . Def. 1. Fremdwort, Fremdwortvorsilbe für selbst. 2. Kurzwort für Automobil.

Lemma Fahrt : POS NOUN . Def. [Einheit] Die Fahrt war ein Wein- und Volumenmass in Freiburg in der Schweiz.

Lemma Fahrt : POS NOUN . Def. [Bergbau] Als Fahrt bezeichnet der Bergmann alle unter Tage eingesetzten Leitern. Er entstammt der Bergmannssprache und geht auf das Verb fahren zurück, das in seiner ursprünglichen Bedeutung jede Art der Fortbewegung bezeichnete. Landschaftlich ist auch die Bezeichnung Fahrte gebräuchlich.

Lemma Fahrt : POS NOUN . Def. [Jugendbewegung] Wandervögel, Pfadfinder in Deutschland und Bündische sagen, sie gehen auf Fahrt, wenn sie bewusst durch die Natur fahren oder wandern. Diese Tradition stammt aus der Jugendbewegung.

Digital Dictionary of the German Language – BBAW Lexical Resources, FCS v2.0 View

Auto fahren : VERB. sich (im Straßenverkehr) mithilfe eines Personenkraftwagens fortbewegen; ein Personenkraftfahrzeug führen, bedienen, lenken

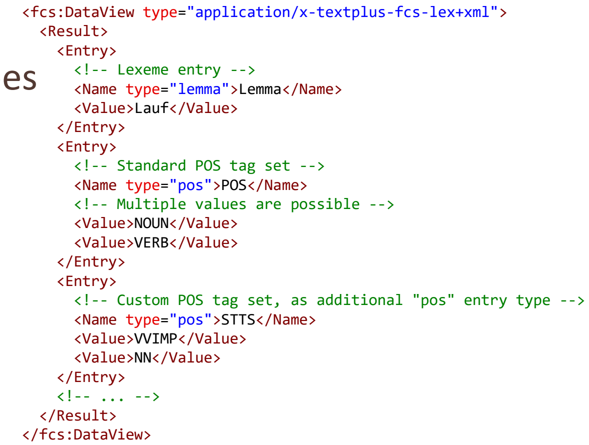
Fahrt : NOUN. ⟨in Fahrt sein, kommen, geraten⟩ / ⟨jmdn. in Fahrt bringen⟩

Fahrt : NOUN. [Seemannssprache] ⟨Kleine, Große Fahrt⟩ Bereich, für den ein Schiff oder ein Schiffsführer zugelassen ist

Fahrt : NOUN. ⟨jmd. hat freie Fahrt⟩ kann ungehindert tun, was er will

Possible Additional Fields for Complex Tabular Data View

- Structured result presentation, in discussion: „key-value pairs“
- Aim: easy conversion of potential complex formats into general flat structure
- Requires:
  - Recommendation for *required* and *optional* information types
  - *Normative* list of keys and value *formats*
- Attributes:
  - Lemma, Pos, Definition, SenseRef, *nym
  - *Examples, Baseform, Hyphenation, Decomposition*
  - *Cooccurrences, Frequency*
  - *Provenience, Etymology/Word-History*



```
<fcs:DataView type="application/x-textplus-fcs-lex+xml">
  <Result>
    <Entry>
      <!-- Lexeme entry -->
      <Name type="lemma">Lemma</Name>
      <Value>Lauf</Value>
    </Entry>
    <Entry>
      <!-- Standard POS tag set -->
      <Name type="pos">POS</Name>
      <!-- Multiple values are possible -->
      <Value>NOUN</Value>
      <Value>VERB</Value>
    </Entry>
    <Entry>
      <!-- Custom POS tag set, as additional "pos" entry type -->
      <Name type="pos">STTS</Name>
      <Value>VVIMP</Value>
      <Value>NN</Value>
    </Entry>
    <!-- ... -->
  </Result>
</fcs:DataView>
```

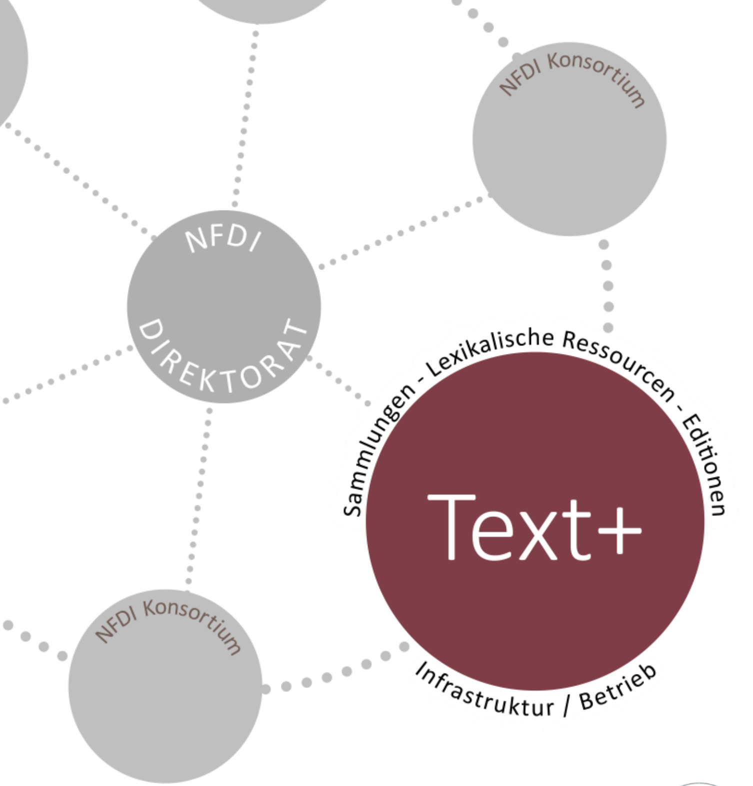
Next Steps and Future Work

Where are we at?

- First draft (v0.1) published
  - LexCQL query language & LexHits data view
  - Demo implementations (Text+ FCS Aggregator + Endpoints)
- Text+: 50 resources from 6 institutes (2 data domains)

Planned

- Implementation Guide
- Tabular Key-Value Data View
- Endpoint Tester for (Lex)CQL Conformance Levels



eLex 2023

Thank you for your attention!

Erik Körner, Saxon Academy of Sciences and Humanities in Leipzig, koerner@saw-leipzig.de

text-plus.org / office@text-plus.org

LexFCS Specification: https://doi.org/10.5281/zenodo.7849753
Text+ LexFCS Aggregator: fcs.text-plus.org
Sources: https://gitlab.gwdg.de/textplus/ag-fcs-lex-fcs-aggregator

Sächsische Akademie der Wissenschaften zu Leipzig

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

nfdi Nationale Forschungsdaten Infrastruktur

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Text+ is a consortium of the nationwide initiative to establish a National Research Data Infrastructure (NFDI) and is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370.