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

Thomas Eckart[1], Axel Herold[2], Erik Körner[1], Frank Wiegand[2]

[1] Saxon Academy of Sciences and Humanities in Leipzig, Leipzig, Germany
[2] Berlin-Brandenburg Academy of Sciences and Humanities, Berlin, Germany
E-mail: {eckart,koerner}@saw-leipzig.de, {herold,wiegand}@bbaw.de

Abstract

The landscape of digital lexical resources is often characterized by dedicated local portals and proprietary interfaces as primary access points for scholars and the interested public. In addition, legal and technical restrictions are potential issues that can make it difficult to efficiently query and use these valuable resources. The research data consortium Text+ develops solutions for the storage and provision of digital language resources which are then provided in the context of the unified cross-domain German research data infrastructure NFDI. The specific topic of accessing lexical resources in a diverse and heterogenous setting with a variety of participating institutions and established technical solutions is met with the development of the federated search and query framework LexFCS. The LexFCS extends the established CLARIN Federated Content Search (FCS) that already allows accessing spatially distributed text corpora using a common specification of technical interfaces, data formats, and query languages. This paper describes the current state of development of the LexFCS, gives an insight into its technical details, and provides an outlook on its future development.

Keywords: lexical resources; federated content search; Text+; information retrieval

1. Introduction

The Text+ consortium[1] works on the utilization of text- and language-based research data in a distributed research environment. It is part of Germany's National Research Data Infrastructure (NFDI[2]) which aims to make research data available for scientific usage, support their interlinkage, and their long-term preservation. Consortia from various research areas are participating in the NFDI and work on establishing an interdisciplinary network of data and services based on common standards and the FAIR principles (findability, accessibility, interoperability, and reusability).

Text+ is organised in three "data domains". The data domain "lexical resources" deals with all kinds of lexical resources, including dictionaries, encyclopedias, normative data, terminological databases, ontologies etc. Many of the largest German providers of such resources are members of the consortium. The data domain is structured in three thematic clusters with varying focuses (see figure 1).

One salient goal of the data domain is the integration of lexical data in a decentralized dictionary platform. Due to the heterogeneous nature of available resources, formats, levels

[1] https://www.text-plus.org/en
[2] https://www.nfdi.de/?lang=en

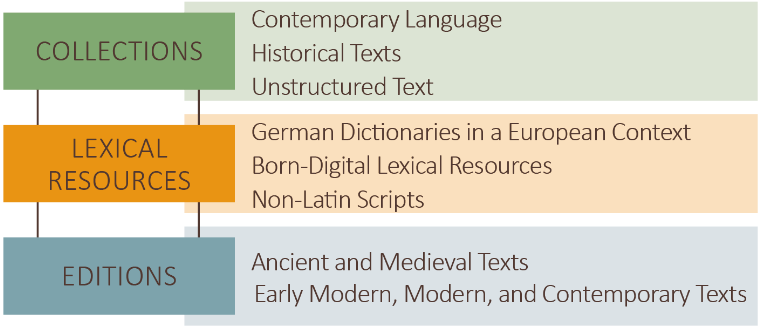

COLLECTIONS
Contemporary Language
Historical Texts
Unstructured Text
LEXICAL RESOURCES
German Dictionaries in a European Context
Born-Digital Lexical Resources
Non-Latin Scripts
EDITIONS
Ancient and Medieval Texts
Early Modern, Modern, and Contemporary Texts

Figure 1: Data domains and their thematic clusters in the Text+ project

of annotation, and technical architectures in use, the implementation will follow a federated approach based on common protocols and formats. Query and retrieval of lexical data on this platform is based on the protocol of the CLARIN Federated Content Search (FCS[3]). The CLARIN FCS is an established framework that already allows accessing spatially distributed text corpora using a common specification of technical interfaces, data formats, and query languages. This framework was developed in the European CLARIN[4] (*Common Language Resources and Technology Infrastructure*) project (Váradi et al., 2008). Its current specification is the basis for adding additional features that support the querying and retrieval of lexical records in the same distributed research environment.

The paper is structured as follows: section 2 gives an overview about related work in the field of providing lexical resources in a distributed research environment. Section 3 describes the state and amount of available lexical resources in the Text+ project. Section 4 outlines the general characteristics of the Federated Content Search, followed by section 5 that describes all extensions made to address specifics of lexical resources. Finally, section 6 presents the conclusion and highlights other current work and plans for further improvements.

2. Related Work

To make different electronic lexical resources available in one place and to allow them to be browsed and queried in a unified way has been a longstanding endeavour for years. Often, such projects were organisationally restricted to single institutions such as is the case with the Trier "Wörterbuchnetz",[5] a growing collection of mainly historical and dialectal dictionaries on the German language. Similar projects can be found across the world for different languages. Another early attempt in this regard is the interconnection of wordnets in different languages as pursued by the Global WordNet Association.[6] However, most of these attempts were focused on lexical resources that are structurally and conceptionally very similar.

With the advent of the creation of common research infrastructures on national and international levels and a strong focus on FAIR data, more general initiatives have tackled

[3] https://www.clarin.eu/content/content-search
[4] https://www.clarin.eu/
[5] https://woerterbuchnetz.de/
[6] http://globalwordnet.org/

the problem of unifying the ways to access and exploit lexical data, such as the ERICs (*European Research Infrastructure Consortia*) CLARIN[7] and DARIAH[8] (*Digital Research Infrastructure for the Arts and Humanities*) together with their national sub-projects as well as the *ELEXIS*[9] project among a range of smaller initiatives.

The four FAIR principles were not targeted equally for lexical data by the early infrastructures. CLARIN and DARIAH focused their efforts especially on *findability* and *accessibility*, resulting in an elaborate metadata ecosystem (e. g. based on CLARIN's *component metadata infrastructure, CMDI*[10]), and a group of distributed certified data centers to operate repositories that host the actual data. The ELEXIS project on the other hand focused more strongly on *interoperability* and *reusability* in terms of the computational exploitation of lexical data. It had a strong influence on the development of the OntoLex/Lemon[11] model for the representation of lexical data (McCrae et al., 2017), and on the ISO 24613 family of standards on the *lexical markup framework*[12] (LMF).

Other initiatives such as the TEI[13] (Text Encoding Initiative) have also worked on the standardization of dictionary and lexicon mark-up since the 1980s. In this context, the focus is often but not exclusively directed on the faithful representation of digitized print dictionaries. Work on the refinement of the TEI guidelines for the encoding of lexical data has recently been promoted by DARIAH and ELEXIS as well as by individual scholars, most notably in the TEI Lex-0[14] initiative.

Orthogonal to the approaches described above, there are also more communal attempts to the creation of lexical resources which are not exclusively run by academic participants. The most important projects in this regard are Wiktionary,[15] DBPedia,[16] and Wikidata.[17] While Wiktionary is essentially a community-driven multilingual dictionary based on (highly formalized) wiki syntax, DBPedia and Wikidata aim at automatically extracting strictly formalized information (though not restricted to lexical information) from sources like Wiktionary and (foremost) Wikipedia and at providing the extracted knowledge in the linked open data (LOD) paradigm, e. g. in the form of RDF serializations.

3. Lexical Resources in Text+

The diversity of (technical) data representation in lexical resources that was outlined above is also reflected in the actual data the participating institutions contribute to the Text+ project and which they have contributed to earlier projects such as CLARIN. Representation formats range from generic and customised TEI/XML serializations to legacy XML formats to table-like serializations (in the cases e. g. of lemma lists and frequency information) to geographic information captured in images of maps and to many more formats. This

[7] https://www.clarin.eu/

[8] https://www.dariah.eu/

[9] https://elex.is/

[10] https://www.clarin.eu/content/component-metadata

[11] https://www.w3.org/2019/09/lexicog/

[12] https://www.iso.org/standard/68516.html

[13] https://www.tei-c.org/

[14] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[15] https://www.wiktionary.org/

[16] https://www.dbpedia.org/

[17] https://www.wikidata.org/

heterogeneity makes a unified representation of the data both for retrieval and presentation a challenging task.

The set of data categories for a given lexical resource is typically specific to this resource. It may range from very broad and general categories (e. g. headword, definition) or mostly uncontroversial grammatical features of the headword (e. g. its part-of-speech) to highly specific linguistic properties that only occur in certain types of dictionaries (e. g. cognates, lexical inheritance relations), or to properties that are strongly bound to certain linguistic theories (e. g. different notations of collocation, or the treatment of homonymy/homography).

Moreover, participating institutions typically have operated local systems for serving, updating, and querying their resources and thus created specific environments for the maintenance and exploitation of their resources. These environments need also not necessarily be technically interoperable per se across different institutions. They may rely on different underlying data formats, different query languages, and different protocols for communicating with their server instances.

Given the situation described above, three principal strategies for harmonising the lexical data and the access to the data can be considered:

1. converting all lexical data into a single common format and using generic software to access the data;
2. explicitly annotating all data categories in the different formats and using generic software that operates on the annotations to access the data;
3. not changing the data or the existing infrastructure and using conversion mechanisms to relay standardised queries to the existing infrastructure and possibly also converting the query results to a specific exchange format.

Converting all lexical resources into genuine triple representations (e. g. in the form of an RDF serialization) based on an agreed-upon predicate inventory would be the easiest way to achieve a unification in case one. The generic infrastructural pillar would then be a triple store. This shifts the computational burden to the representation of the results. These will then have to be transformed into a human readable form in the context of a general research infrastructure. In this paper we do not report on preliminary work we have done in this direction.

Following case two, all data categories would have to be marked up in the lexical data. Depending on the granularity of the categories this can lead to tedious manual or automatic annotation work on all lexical data sources. The generic software would still have to be able to work on different serializations (e. g. XML vs. JSON vs. proprietary formats such as relational databases).

The scenario in case three has the advantage that the lexical data does not have to be modified or adapted. This can be essential when the data is also used in other workflows such as is the case for ongoing editions that rely on a software stack of their own. What needs to be implemented, though, is an on-the-fly transformation of incoming standardised queries into the resource specific query language. Note that this transformation might not be guaranteed to be lossless when the source query language has greater expressive power than the target query language.

In the following sections we describe our approach with respect to the third case.

4. Federated Content Search Infrastructure

The CLARIN Federated Content Search (FCS) is an established federated search engine that is specified, developed, and maintained in the context of the European CLARIN project. CLARIN works on an interoperable, integrated research online environment for the support of researchers in the humanities and social sciences. CLARIN is characterised by many participating institutions (so called *CLARIN centres*) that provide linguistic resources for a variety of research communities. These centres agree on and adhere to general requirements on how to provide data, tools, and services and work on an integrated research environment where those resources are linked by and accessed via common data formats and technical interfaces.

In this context, the CLARIN FCS' original focus is to give access to text corpora in this environment. It allows querying distributed corpora by using a standardised RESTful protocol and data formats (Stehouwer et al., 2012) that are based on the *Search/Retrieval via URL* protocol (specified by the Library of Congress, Morgan (2004)) and the *searchRetrieve* protocol (OASIS, 2013), specified by the open standards consortium OASIS (Organization for the Advancement of Structured Information Standards). The protocol allows to query data stored in online available data 'endpoints' via three operations of which the following two are relevant here:

- Operation `Explain` to identify capabilities (like supported query language(s), query vocabulary, and data formats) and available resources that a specific endpoint provides.
- Operation `SearchRetrieve` to query (a subset of) those resources at a specific endpoint.

Based on these operations, a client – including central aggregators or search portals – can query a single endpoint, or multiple endpoints in parallel. Each endpoint functions as a "proxy" for the local technical infrastructure at a specific institution in the FCS infrastructure and is typically hosted by the individual institution itself (see figure 2).

The SRU/searchRetrieve protocol includes means to be easily adapted to new requirements. The CLARIN FCS makes use of this mechanism and extends the protocol with a focus on accessing text corpora. These text corpora can be queried based on their fulltext representation or by addressing a variety of linguistic annotation layers, including part-of-speech, word baseforms, or phonetic transcriptions. For this, a dedicated corpus query language FCS-QL (inspired by the popular CQP[18]) and data representation formats were defined as key components of the protocol.

The CLARIN FCS acknowledges the problem of heterogeneity in a distributed research environment, where access to data can vary in aspects like the data format used, storage solution, query language and more. By agreeing on a lightweight retrieval protocol and simple default data formats (so-called *DataViews*) those distributed resources can be

[18] https://cwb.sourceforge.io/files/CQP_Manual/

made available to end users via easy-to-use Web interfaces (like the FCS aggregator[19]). In this sense, the FCS does not provide a feature-complete replacement of specific search interfaces, but offers a simple way to get an overview of available resources that can also be accessed by the specialised applications at the hosting institution, if needed.

Despite its original focus on text corpora, support for requests and retrieval of lexical entries in the FCS has long been discussed and is currently implemented in an iterative work process coordinated between Text+ and CLARIN's FCS taskforce. This seemed to be especially reasonable as many German Text+ participants already participate in the current CLARIN FCS infrastructure and therefore have experience in creating and maintaining compatible endpoints.

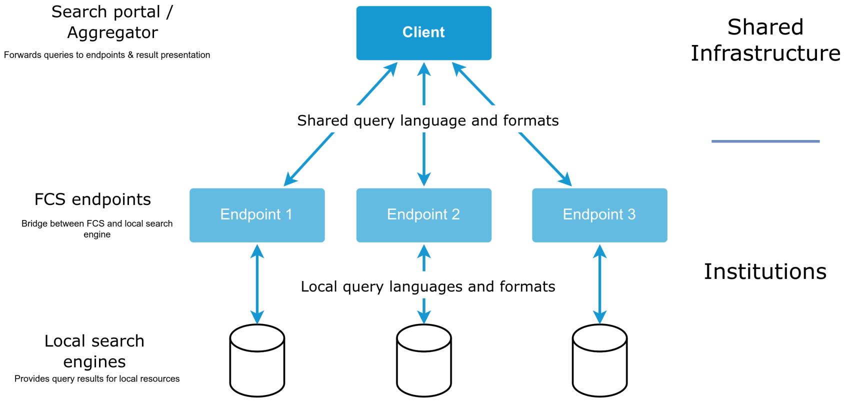

Search portal / Aggregator
Forwards queries to endpoints & result presentation
Client
Shared Infrastructure
Shared query language and formats
FCS endpoints
Bridge between FCS and local search engine
Endpoint 1
Endpoint 2
Endpoint 3
Local query languages and formats
Institutions
Local search engines
Provides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

Figure 2: The general FCS architecture

5. FCS Specification Extension for Lexical Resources

The FCS specification (Schonefeld et al., 2014) in its latest version 2.0 describes two search modes:

- **Basic Search** (mandatory) is the minimum requirement to participate in the FCS infrastructure. It specifies a minimal query language for fulltext search and a simple *HITS* (Generic Hits) DataView (basic highlighting of query matches and Keyword-in-Context (KWIC) visualization) for results. This search mode allows to integrate any resource that has some form of textual representation.
- **Advanced Search** (optional) is used for searching in annotated text data with one or more annotation layers. The specification describes six types of layers (text, lemma, part-of-speech, and different forms of normalisation and transcription) for – potentially complex – queries using FCS-QL, a CQP-like query language. The result serialisation is the *ADV* (Advanced) DataView to support structured information in annotation layers. Annotations are (character) streams with offsets which also allows for e. g. audio transcriptions. This search capability is primarily focused on text corpora and similar resources.

We extend the core FCS specification in terms of *announcing*, *querying* and *retrieving* lexical resources while we ensure to seamlessly integrate and remain compatible with the

[19] https://contentsearch.clarin.eu

overall FCS architecture. This allows to reuse features such as access control for restricted resources, automatic configuration of clients, and the overall registration of endpoints within the FCS system (see figure 3). We also adapt existing search interfaces to support users in the process of creating lexical queries and dealing with the results offered (see figure 4). The specification extension for lexical resources introduces the **LEX** search capability, and entails:

- Specifying the query language (see section 5.1) which is a "CQL Context Set"[20] of the Contextual Query Language[21] (standardized by the US Library of Congress) dedicated to querying lexical entries. Its specification includes agreements on accessible fields of information (like part-of-speech, definitions, (semantically) related entries etc.) for a lexeme, and how to combine them to complex queries. This is especially challenging due to the inherently hierarchical structure of lexical data.
- Specifying common data formats for a unified result presentation (see section 5.2). On the basic level, this is achieved by a mandatory KWIC representation that allows annotating information types inline and by an advanced tabular-representation of all fields in a key-value-style. It is clearly understood that in most cases these representations can only provide a simplified view on the data. It is therefore endorsed to provide records in their complex native representation as well with examples being different TEI dialects including TEI Lex-0,[22] OntoLex/Lemon,[23] and other formats.

For the current draft of the LexFCS proposal for extending the core FCS specification with regards to lexical resources refer to Körner et al. (2023). The document is still under heavy development and subject to change.

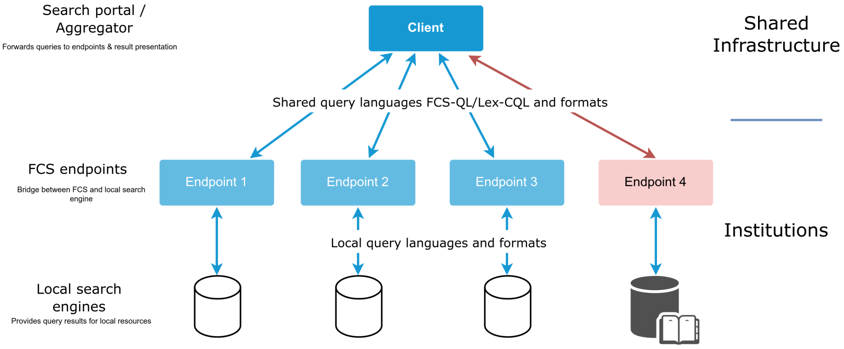

Search portal / Aggregator
Forwards queries to endpoints & result presentation
Client
Shared Infrastructure
Shared query languages FCS-QL/Lex-CQL and formats
FCS endpoints
Bridge between FCS and local search engine
Endpoint 1
Endpoint 2
Endpoint 3
Endpoint 4
Institutions
Local query languages and formats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

Figure 3: The FCS architecture extended to incorporate endpoints for lexical resources

5.1 Query Language – LexCQL

We propose **LexCQL** as the main query language – a subset of the CQL[24] which is customized for searching through fields of lexical resources. In contrast to text corpora that

[20] https://www.loc.gov/standards/sru/cql/contextSets/theCqlContextSet.html

[21] https://www.loc.gov/standards/sru/cql/

[22] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[23] https://www.w3.org/2019/09/lexicog/

[24] Please note that in this paper the term "CQL" is always referring to the *Contextual Query Language* of the Library of Congress and should not be confused with the Sketch Engine's *Corpus Query Language*.

are subdivided into sentences, paragraphs, documents etc., with their various annotation layers, lexical resources are often organised around single lexical entries with specific information that is frequently represented in the form of property-value pairs. Even though lexical entries can be nested, the LexCQL initially focuses on flat entries only.

A typical minimal set of information available in many Text+ resources contains the following searchable information types (or “indexes” in CQL):

- **lemma**: Lemma or article name,
- **pos**: Part-of-speech; it is encouraged to support – in addition to potentially more specific tagsets – the *Universal POS tags* of the Universal Dependencies project,[25]
- **def**: Definition or description as fulltext string,
- **xr$synonymy**, **xr$hyponymy**, ...: Semantic relations as fulltext strings; analogous to the TEI Lex-0 types,[26]
- **senseRef** (draft): ID/URI refering to external authority files or lexical databases, like Princeton WordNet, GermaNet, GND, or WikiData.

In the current specification draft, only the relation “`=`” is defined to separate queried field and value. In general, endpoints should be lenient when processing queries to improve usability and recall of results. This might include to implicitly handle spelling variants, to use normalisation procedures for historic word forms, or to support partial matches for full text fields like definitions. The CQL relation modifier “`/exact`” should be used and supported when searching for an exact string match.

For more complex queries, Boolean operators[27] such as `AND`, `OR` and `NOT` can be used and structured via parentheses if necessary. As specified by CQL, strings containing whitespaces or special characters require quoting using doubles quotes (") which are optional otherwise. However, we suggest using quotes for better readability.

Examples:

```
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"
```

5.2 Data Format for Results

The result formats currently supported by the FCS (HITS and ADV “DataViews”) are insufficient for the structure of lexical resources like dictionaries, encyclopedias, wordnets, or ontologies. The LexFCS specification proposes two additional formats.

The DataView **LexHITS**, an extension of the basic HITS DataView allows endpoints to optionally annotate information (like lemma, part-of-speech, and the record’s definition,

[25] https://universaldependencies.org/u/pos/

[26] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#crossref_typology

[27] https://www.loc.gov/standards/sru/cql/contextSets/theCqlContextSet.html#booleans

explanation or description) in a fulltext representation. This allows endpoints to reuse the mandatory KWIC format of the HITS DataView to present a simple representation of the entry but augments the results with more information. If this is not feasible, endpoints can gracefully fall back to a plain text result. The information types to be annotated are intentionally kept similar to the three most basic LexCQL fields (*lemma*, *pos*, and *def*) to emphasize the relation between queried fields and result presentation. To remain compatible with the HITS DataView the search hits marker (`<hits:Hit>`) is reused and extended by the XML attribute `@kind`. For a specific example of its use in a result's presentation, see figure 4.

Example of HITS DataView with Lex annotations extension (highlighted in red):

```
<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.

For a more structured presentation of results, an optional DataView is currently discussed that allows providing lexical information as key-value pairs. This format aims at an easy conversion of potentially complex formats into a more general – however simplified – flat structure. As it focuses on a shallow representation, nested entries with sub-structures will need to be flattened into their own entries for search and retrieval. Discussions are ongoing to specify a set of recommendations on required and optional information types and a normative list of keys and value formats.

Example of a potential tabular key-value DataView:

```
<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>
```

It is also suggested that resources are made available in their native representation, e. g., in various TEI dialects including TEI Lex-0, OntoLex/Lemon, and other formats in custom DataViews. If necessary, stylesheets, e. g. XSL(T), can be used as a generic way to transform TEI-based or XML-serialized RDF formats into a uniform presentation.

5.3 Search Portal / User Interface Prototype

The prototypical LexFCS search portal implementation is already available as a basis for further discussion and refinements.[28] It provides access to endpoints maintained by several lexical resource providers of Text+. A first stable version of the specification and an improved user interface implementation is expected until end of 2023. As a means of technical integration, the LexFCS aggregator provides an OpenAPI-compliant specification of its RESTful API.[29]

5.4 Software and Software Libraries

The source code of all infrastructural components is provided using open-source licenses. This includes the central search portal,[30] parsers and validators for LexCQL in various programming languages,[31] and specification and documentation artifacts.

6. Next Steps and Future Work

All mentioned constituents of the architecture are actively worked on and are incrementally developed. Throughout specification and implementation, feedback is provided by interested parties, particularly from but not limited to the Text+ and CLARIN projects. With a first public release in the coming months – based on the current demonstrator –, the availability and visibility of various lexical resources will be improved, including some that were not easily accessible or even unknown to the general public until now. Future work will therefore be focused on finalising the specification for the lexical search functionality. This includes the broader dissemination of the specification and providing reference implementations by different parties.

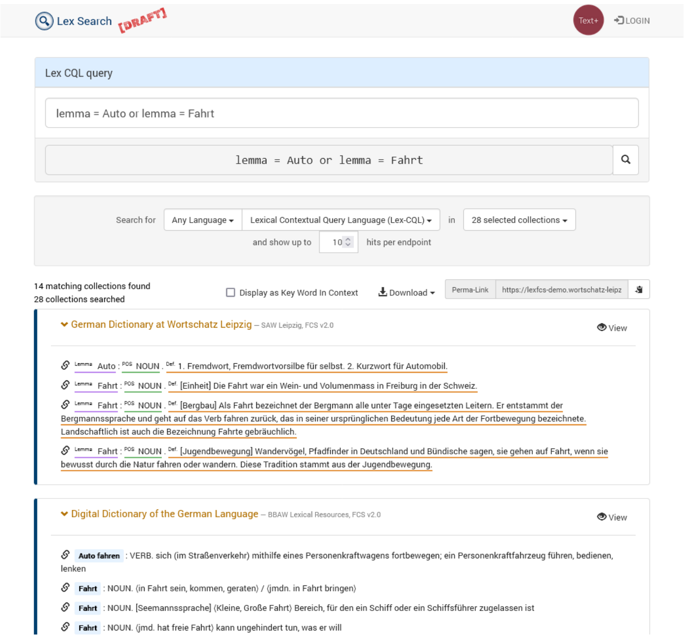
One general question that has become salient during the previous work is the problem of accessing restricted resources. Those restrictions – often because of legal obligations with

[28] https://hdl.handle.net/11022/0000-0007-FBF2-D

[29] https://hdl.handle.net/11022/0000-0007-FBF2-D?urlappend=%2Fopenapi.json

[30] https://gitlab.gwdg.de/textplus/ag-fcs-lex-fcs-aggregator

[31] https://gitlab.gwdg.de/textplus/



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

Perma-Link https://lexfcs-demo.wortschatz-leipz

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

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

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

Figure 4: Screenshot of the current frontend demonstrator

publishers that prevent public access – are adressed by extending the current specification on (a) how to notify users about possible restrictions on resources, (b) how to present possibly restricted results to an end-user, and (c) how to formalise the access modalities of those resources. Scenarios might include only authenticated users being able to view results as well as providing meta information about possible hits with users being only able to view actual results at the institute or publisher in question. Using the established *Authentication & Authorization Infrastructure* (AAI) for federated authentication mechanisms with *SAML/Shibboleth* (Needleman, 2004) is currently being worked on and specifications as well as working prototypes are planned during this year in the context of the overall development of the CLARIN FCS.

A major problem of federated search systems is the absence of a global result ranking. Due to the distributed nature of the FCS, each endpoint decides how to rank its results. Those criteria are often not comparable because of differences in local retrieval systems or even the nature of the resources themselves. Results in the aggregator are therefore only grouped by resource and providing endpoint, but not in a joint representation. Using collection or provider based ranking approaches (Shokouhi & Si, 2011), result preference based on specificity of records regarding a concrete query, or other standard information retrieval methods might be a sensible approach.

Records containing lexical information referring to identical lemmas from different providers are also an issue that can significantly reduce the usability for end users. It is planned to evaluate the usage of external references to semantic wordnets – like Princetown WordNet (Fellbaum, 1998) or GermaNet (Hamp & Feldweg, 1997) –, authority files – like the *Integrated Authority File* of the German National Library (DNB, 2023) –, or other knowledge bases – like Wikidata (Vrandečić & Krötzsch, 2014) or Wiktionary – to allow a sense-based combined representation of information from different data providers.

As the software is already publicly available, third parties that wish to make their lexical data accessible over the FCS infrastructure can already set up endpoints for the aggregator.

They can also deploy a self-contained instance of the FCS including their own aggregator. Based on the specification, independent software solutions can also be developed, e. g. based on the TEI publisher.[32] However, we are not currently planning to provide software beyond the reference implementations.

7. Acknowledgements

This publication 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, the authors would like to thank all members of the Text+ data domain *Lexical resources* for their continuous work.

8. References

DNB (2023). The Integrated Authority File (GND). https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html. Accessed: 2023-04-14.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books. URL https://mitpress.mit.edu/9780262561167/.

Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. URL https://aclanthology.org/W97-0802.

Körner, E., Eckart, T., Herold, A., Wiegand, F., Michaelis, F., Bremm, M., Cotgrove, L., Trippel, T. & Rau, F. (2023). *Federated Content Search for Lexical Resources (LexFCS): Specification*. URL https://doi.org/10.5281/zenodo.7849753.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In *Proceedings of eLex 2017*. pp. 587–597.

Morgan, E.L. (2004). An Introduction to the Search/Retrieve URL Service (SRU). *Ariadne*, 40. URL http://www.ariadne.ac.uk/issue/40/morgan/.

Needleman, M. (2004). The Shibboleth Authentication/Authorization System. *Serials Review*, 30(3), pp. 252–253. URL https://www.sciencedirect.com/science/article/pii/S0098791304000978.

OASIS (2013). *searchRetrieve: Part 0*. Organization for the Advancement of Structured Information Standards. URL http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/searchRetrieve-v1.0-part0-overview.html.

Schonefeld, O., Eckart, T., Kisler, T., Draxler, C., Zimmer, K., Ďurčo, M., Panchenko, Y., Hedeland, H., Blessing, A. & Shkaravska, O. (2014). *CLARIN Federated Content Search (CLARIN-FCS) – Core Specification*. URL https://www.clarin.eu/content/federated-content-search-core-specification.

Shokouhi, M. & Si, L. (2011). *Federated Search*. Federated search edition. URL https://www.microsoft.com/en-us/research/publication/federated-search/.

Stehouwer, H., Durco, M., Auer, E. & Broeder, D. (2012). Federated Search: Towards a Common Search Infrastructure. In *Proceedings of the Eighth International Conference on*

[32] https://teipublisher.com/index.html

Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA), pp. 3255–3259. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/524_Paper.pdf.

Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M. & Koskenniemi, K. (2008). CLARIN: Common Language Resources and Technology Infrastructure. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/317_paper.pdf.

Vrandečić, D. & Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10), p. 78–85. URL https://doi.org/10.1145/2629489.