

# **An Enhanced Federated Content Search Infrastructure for the Humanities and Social Sciences**

**Erik Körner, Thomas Eckart, Felix Helfer, Uwe Kretschmer**

Saxon Academy of Sciences and Humanities in Leipzig

Leipzig, Germany

`{koerner,eckart,helfer,kretschmer}@saw-leipzig.de`

## **Abstract**

The general idea and implementation of a federated search infrastructure component that allows querying both full-text resources and their linguistic annotations is a prominent part of the CLARIN project and is closely interconnected with the other components of its decentralised European-scale research data infrastructure. Since its beginnings, the Federated Content Search (FCS) has been continuously improved and by now fulfils its original goals that were formulated more than 12 years ago. During the last years, development of the FCS has accelerated massively with newly formulated application scenarios, newly opened up user groups and newly developed tools and user interfaces. This paper gives a summary of the developments of recent years and the topics that are currently being worked on. In addition to the further development of existing modules, this includes in particular the consideration and implementation of new requirements reflecting a rapidly evolving research infrastructure landscape.

## **1 Introduction to the Federated Content Search**

The Federated Content Search (FCS) is a system for facilitating the discovery and retrieval of linguistic resources distributed across various data providers. The key idea is to enable users to search through different resource types (including text corpora and dictionaries) stored in multiple repositories in parallel through easy-to-use interfaces. Instead of requiring users to search each individual resource separately, the FCS aggregates results from these resources and provides a search capability that allows users to access and explore a wide range of linguistic content seamlessly. The FCS was developed in the context of the European CLARIN project. Today, it offers access to more than 500 resources<sup>1</sup> provided by scientific institutions from twelve European countries.

The general idea of the Federated Content Search was outlined in 2012 (Stehouwer et al., 2012) and mentioned a number of topics as central characteristics. They include searching in research data content – as opposed to searching metadata records –, support of distributed resources, use of a standard protocol (SRU/CQL, OASIS, 2013), and consideration of possible future extensions. The FCS as it is today represents an easy-to-use solution to execute complex linguistic queries on large distributed data sets. It does not replace the powerful corpus query engines already in use, but provides a lightweight interface to run parallel queries on them. The connection between the FCS and the respective local search engine is made via so-called “FCS endpoints” which map FCS queries to the locally supported query language and convert local data formats into FCS-compatible formats.

In the aforementioned publication, various topics were described as possible future priorities, including the provision of bindings for popular data indexing systems, integration with other CLARIN services, support for access-restricted resources via suitable control mechanisms, and the use of ISOcat (Kemps-Snijders et al., 2009) as a semantic integration layer between the common infrastructure and specific endpoint implementations. Version 1.0 of the FCS specification (Schonefeld et al., 2014) two years after

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

<sup>1</sup>FCS Aggregator, <https://contentsearch.clarin.eu/>, as of April 2025

the original paper incorporated large parts of these ideas and was superseded by version 2.0 (van Uytvanck et al., 2017) in 2017, which introduced a new query language *FCS-QL* and a new data view to enhance access to annotated text corpora.

The vast majority of the initially planned ideas have now been implemented and, in most cases, continuously improved and expanded in several iterations. Of the original ideas, only two topics are currently still unfinished, which have significantly different probabilities for implementation: the support of access mechanisms for querying protected resources – which is currently in progress – and the use of ISOcat as an integration layer which is no longer part of the development goals<sup>2</sup>.

The FCS continues to be a central component of the CLARIN infrastructure and is a prominent part of its current work plan, but has also established itself in institutions and projects beyond this. A prominent example is the establishment of the FCS as the central search component in the German infrastructure consortium *Text+*<sup>3</sup>, which has already significantly increased the number of available resources. In the same context, new user interfaces are currently being developed to improve usability and user-friendliness and new usage scenarios are increasingly being supported that benefit from the flexible extensibility of the FCS standard.

## 2 New Requirements in a Changing Resources Landscape

The original concept of FCS focussed on resources that are mostly available as tokenized continuous text (such as plain text documents, text corpora, or text collections) with optional token-based linguistic annotations that are freely accessible in a decentralized infrastructure.

These key aspects are still valid for most of the resources provided. However, it is becoming increasingly clear that this does not adequately reflect the diversity of linguistic resources or resource types and that the all too common scenario of trying to access protected resources (due to publisher restrictions, copyright or personal rights, etc.) limits the availability of text resources that are highly relevant for research. Research projects are often oriented towards and organised based on specific resource types, with sometimes significantly varying technical requirements and traditions. In the German *Text+* project, for example, this is illustrated by its organisation into three data domains “Collections”, “Lexical Resources” and “Editions”. During years of project work, it became clear that the FCS in its current form not always sufficiently covers their requirements. This becomes especially obvious for lexical resources (such as dictionaries, word lists, lexical-semantic networks), which differ greatly from full-text-oriented resources due to their sometimes complex internal structure and which also require a different set of queryable types of information.

The increasing importance of central knowledge bases – such as Wikidata (Vrandečić & Krötzsch, 2014), VIAF<sup>4</sup>, or the Integrated Authority File GND (German National Library (DNB), 2024) – for the discovery of and linkage between digital resources in the humanities research cannot be sufficiently supported within the existing FCS implementation. This support of authority files or referenceable knowledge bases with similar functionality is addressed within the *EntityFCS* specification. The first corpora and dictionaries that make use of this functionality are already available.

In comparison, the support of access-restricted resources via an Authentication & Authorization Infrastructure (AAI) is a long-standing requirement. This includes the option for users to log in via their known identity provider if required and to forward attributes to FCS endpoints that can allow access to resources that are only available to restricted user groups or even individual users. Seamless integration into existing user interfaces is a key aspect in ensuring a high level of usability here. However, developments in recent years have also led to adapted requirements in this area. Especially “reference-only results” – indicating only the presence of a result at a specified location – or support of *derived text formats* (Schöch et al., 2020) are being discussed for cases where the actual content cannot be provided due to contractual or legal reasons.

---

<sup>2</sup>Due to its discontinuation in 2014.

<sup>3</sup>Text+, <https://www.text-plus.org/en>

<sup>4</sup>Virtual International Authority File, <https://viaf.org/>

Another problem that illustrates the diversity of language resources in the context of incomplete standardisation is the inadequate representation of character sets, which often occur in the field of historical languages and which are not always adequately supported by standards such as Unicode<sup>5</sup> or by established and freely available fonts. Especially for a federated search platform, resulting display artifacts pose a major problem for an appropriate representation of results. This problem is addressed by an optional extension of the FCS specification that allows endpoints to provide font information for resources, which can be presented to the user in the search interfaces.

This paper is structured along these new requirements, starting with a brief introduction as to why the FCS is a solid foundation for these types of customisations: the flexible structure of its specification that allows backward-compatible extensions (section 3). In the following sections, these extensions are described in more detail starting with the support of new resource types in the *LexFCS* (section 4) and new kinds of requests in the *EntityFCS* (section 5). The support of queries on access-restricted resources by integrating the FCS into an Authentication and Authorization Infrastructure (AAI) and of appropriate UI representation of resources via external font support is explained in section 6 and section 7. In section 8 an overview is given of different means to improve usability by a revised development and documentation process which significantly facilitates participation in the development process and the provision of own resources and their use by end users. The publication concludes with suggestions for improvement that are currently being evaluated (section 9) or are part of the short- and medium-term time plan.

### 3 Flexibility through the FCS Extension System

Extensibility and backwards compatibility are key factors and important design principles of the FCS. The SRU/CQL protocol (OASIS, 2013) contains a number of supporting mechanisms that were the reason for its choice as the technical basis of the platform. This allows the FCS to adapt to evolving user requirements while maintaining compatibility for endpoints and client libraries.

This flexibility includes extended API calls, support for new data formats (so-called “Data Views”) and new query languages or optional extensions to standardized schemata. All work presented in the following sections ultimately uses these mechanisms. Extended endpoint implementations are – in the vast majority of cases – still accessible to existing client libraries and fit easily into the existing infrastructure.

Figure 1 summarises the overall FCS architecture, highlighting current extensions of its ecosystem.

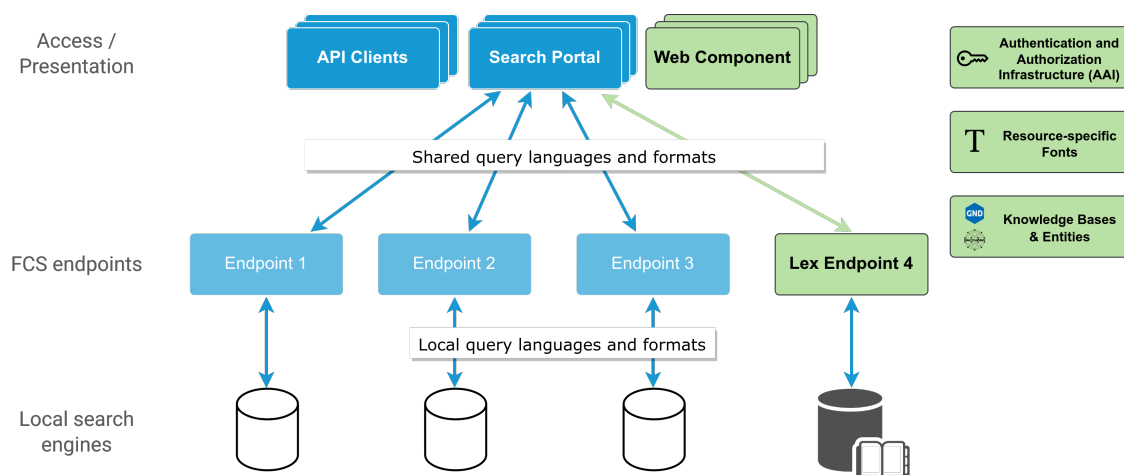


Figure 1: General architecture of the FCS, highlighting currently developed extensions in green

### 4 LexFCS – Lexical Resources in the FCS

The FCS was initially developed to offer a federated content search on simple full texts but later extended for more complex linguistic search patterns on tokenized texts with optional annotation layers. Common to these is that searches are based on “flat” text sequences, making it difficult to map more

<sup>5</sup>Unicode – The World Standard for Text and Emoji, <https://unicode.org>

complex structures into this format. However, this excludes resource types such as lexical resources, including dictionaries, word lists, or semantic wordnets, which typically have a more complex structure, like graphs, feature structures, or simple property value structures.

A first working proposal had been prepared by a dedicated working group in Text+ for a new *LexFCS* extension (Eckart et al., 2023). First prototypes of FCS endpoints and result presentation in FCS clients were developed based on an early and temporary specification. Collected feedback and input from discussions at conferences and workshops were incorporated into the current, revised version of the LexFCS specification (Körner et al., 2024) that proposes the following two major additions, which can only be presented here briefly.

(1) A key-value based data model which structures lexical information in *Entries* that contain typed *Field* elements each having *Values* and supplementary attributes, and which is serialized into a new **Lex Data View**. Each *Entry* represents an individual, self-contained dictionary entry, containing information about a lemma without making assumptions about its type. Fields group properties of a lemma by their information type such as basic information (e.g., word form, spelling variants, identifiers, transcriptions), relations to other entries or external resources (references, citations), morphosyntactic (part-of-speech, segmentation), semantic (like sentiment, synonyms, hyponyms), or frequency-related data as well as more prosaic information (definition, etymology). The actual content is given in *Value* elements and may contain additional meta information about content language, internal or external references, and more.

The internal reference and grouping mechanism is achieved via IDs which allow custom highlighting in the user interface of related elements but can also be used to structure information hierarchically based on given definitions and etymologies.

Regardless of this, the representation of a lexical record in the style of a printed dictionary with additional annotations is now supported with the **LexHITS Data View**. It is a backward-compatible upgrade of the HITS Data View that allows inline annotation of lexical information in full text.

(2) Extending the *Contextual Query Language* (CQL<sup>6</sup>) standardized by the US Library of Congress and the standards organisation OASIS, we propose **LexCQL** as the query language dedicated to querying lexical entries. It defines searchable indexes based on the *Fields* of the aforementioned data model with relations and relation modifiers to allow both simple and complex queries.

LexCQL supports the default search relation = for relaxed and the additional == for strict equality matching. Relation modifiers for case sensitivity, (umlaut) normalization, matching, etc. can be used to refine queries. In addition, the new relation *is* can be used to search based on concept or reference URIs instead of plain value strings, e.g. requesting a noun via its definition according to the Universal Dependencies project (<https://universaldependencies.org/u/pos/NOUN>) instead of highly ambiguous search terms (like “N”). Boolean operators AND, OR, NOT and parentheses allow to combine and build complex queries.

Based on the *Values*’ language information and a language index for the overall *Entry*, queries in multilingual dictionaries are also supported.

Current work has been focused on the standardisation of terminology and to provide a data model and query language to represent and search through potentially complex lexical records while maintaining a high level of flexibility to allow compatibility with many types of lexical resources. The Text+ FCS Aggregator<sup>7</sup> already includes a first implementation of this extension with a few endpoints providing dictionary data. Another way of representing LexFCS results has been implemented in an alternative user interface (cf. Figure 2, more about this search integration in section 8).

Ongoing work will address further technical details to provide improved user interfaces and to offer guidelines for new data providers about available features and on how to integrate their own custom formats into the LexFCS.

<sup>6</sup>Please note that in this paper the term “CQL” is always referring to the *Contextual Query Language*.

<sup>7</sup>LexFCS search in the Text+ FCS Aggregator, <https://fcs.text-plus.org/?queryType=lex>

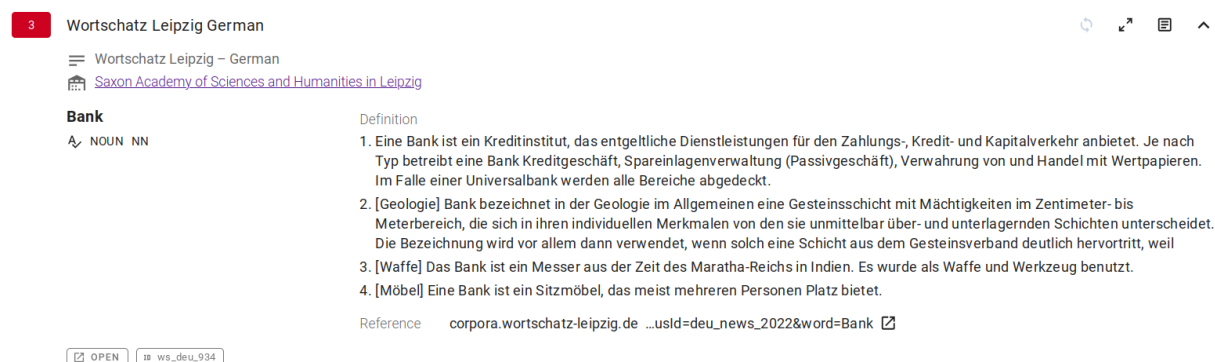



Figure 2: Representing the micro structure of a dictionary entry using the new Lex Data View in a user interface prototype, 

## 5 EntityFCS – Entity-oriented Search

Authority files are important tools to ensure consistency and accuracy in the identification of entities such as persons, locations, organizations, events, and more. Resources that are annotated accordingly allow access to these entities despite language boundaries, potential variations in spelling or transliteration and – even more important – in case of lexical ambiguity the specific intended meaning can be described and referenced. Authority files have therefore the potential to improve data discoverability and interoperability in a distributed and diverse research environment significantly.

In the field of Humanities and Social Sciences, various knowledge bases are increasingly used to enhance existing resources or are considered from the start for born-digital resources. In the context of this section, the term “authority file” is meant in a deliberately broad sense, including all systems that define specific meanings for entities in a field of knowledge and provide persistent means for their identification and global access. As a consequence, not only traditional authority files like the *Virtual International Authority File* VIAF or the *Integrated Authority File* GND (German National Library (DNB), 2024) are considered, but collaborative knowledge bases such as Wikidata (Vrandečić & Krötzsch, 2014) or GeoNames<sup>8</sup> and sense-based lexical databases like GermaNet (Hamp & Feldweg, 1997) and Princeton WordNet (Fellbaum, 1998). The growing importance of Knowledge Graphs for structuring diverse research landscapes and providing access to resources is only conceivable taking such semantic “anchors” into account.

Figure 3 demonstrates the use of two of these knowledge bases in letters to Alexander von Humboldt at the *edition humboldt digital*<sup>9</sup>.

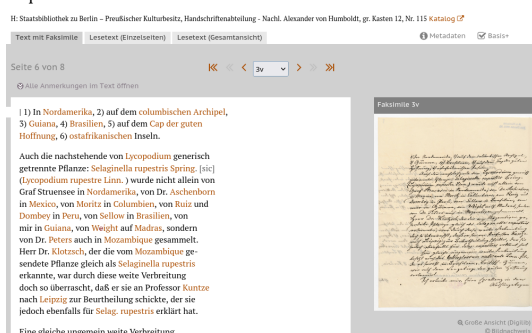
The task of resolving a mention in a text to an appropriate entry in an authority file is called *entity linking* (Hachey et al., 2013). While entity linking can greatly enrich textual resources, the process is often laborious and difficult, especially for linking to a large authority file, where a mention could refer to multiple different entries (e.g. different people with the same name). In addition, previous automated approaches often underperform, especially for non-English languages (see benchmark by Schwarz and Barth, 2024). Hence, data with annotations of this kind are still quite rare. However, with the continued development of new approaches to entity linking, particularly those utilizing the advancement of large language models, such as Qi et al., 2024 and knowledge graphs, such as Ayoola et al., 2022, it is to be expected that data with mentions linked to an authority file entry will become more common over time. Search applications that support these types of annotations can then further help with the accessibility and usability of the referenced data.

The *entity-oriented search* paradigm (Balog, 2018) focuses on retrieval mechanisms on the basis of

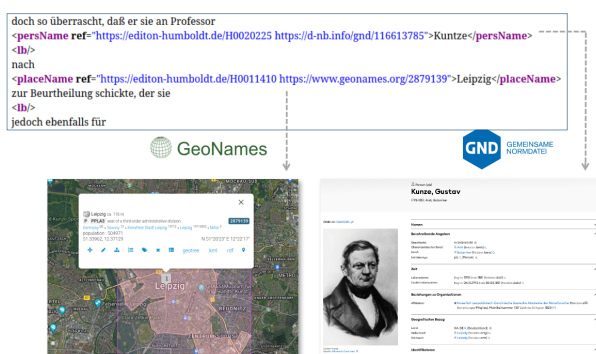
<sup>8</sup>GeoNames, <https://www.geonames.org>

<sup>9</sup>edition humboldt digital, <https://edition-humboldt.de>

Richard Schomburgk an Alexander von Humboldt. Berlin, 9. September 1847



(a) Online presentation, 



(b) TEI encoding of authority file references

Figure 3: Referencing entities in a letter to Alexander von Humboldt using GeoNames and the GND (Project *edition humboldt digital*)

linked entities. However, the FCS in its original specification does not offer any support for this kind of queries. The focus on linguistic annotations with a limited set of annotation layers designed for this dedicated purpose (such as lemma, orthographic transcription and similar) currently supports queries best through the use of the *pos* layer, with which – using suitable vocabulary – at least the information on entity types can be accessed.

The *EntityFCS* extension of the FCS supports search on entity- and sense-annotated full texts as well as lexical resources using global references. As all supported FCS query types have different specifics, there are slightly different characteristics in the respective specification customisation. They both entail modifications to the query language and support of the entity annotation layer in the respective data view:

- **Advanced Search** queries allow searching on the additional layer type “entity” as an optional extension to the existing annotation layers. This is represented as a new layer both in the query language *FCS-QL* and in the serialization in the *Advanced Data View*.
- **Lexical Search** queries allow access to entity references via the field “senseRef” that is supported both in the query language *LexCQL* and in the *Lex Data View*.

Figure 4 shows an example result set in a new web interface (for more details, see section 8), where all textual references to the German city Leipzig are queried based on its GND identifier in a corpus of the project “Letters and Records regarding the Church Politics of Frederick the Wise and John the Constant from 1513 to 1532”<sup>10</sup>. The short description and illustrative image in the blue tooltip are dynamically loaded from the *lobid* API<sup>11</sup>.

The described functionality is already part of the current LexFCS specification (Körner et al., 2024). Its support in the multi-layer advanced search is still in the preparatory stage.

## 6 AAI – Querying Access-restricted Resources

Data protection and copyright laws, licensing, and other reasons often impede general public access to linguistic resources. In order to make restricted content available via the FCS anyway, with interested parties having a better overview of potentially relevant data and data providers having the option to offer at least some kind of access, integration of a decentralized *Authentication and Authorization Infrastructure* (AAI<sup>12</sup>) is required. This enables data providers to check search requests to their endpoints if they

<sup>10</sup>Project homepage, <https://bakfj.saw-leipzig.de/>

<sup>11</sup>Linked Open Data (LOD) services for libraries, <https://lobid.org/index-en>

<sup>12</sup>A brief overview on “What is AAI?”, <https://aai.mpg.de/aai.shtml>

. Johann zugunsten des Beginns der Reform des Dominikanerinnenklosters zu Weida Gelder in Höhe von 78 rheinischen Gulden , 15 Groschen , 2 Pfennigen und 1 Heller ausgelegt hat .

OPEN

ID bakfj:1791

[ 10 ] Die Räte Wolf von Weißenbach

gnd:118716921 entity

Sachsen wegen des **Leipziger** Brief

OPEN

ID bakfj:2216

Sie haben sich deswegen bereits bei Hz . Georg von **Sachsen** beschwert , der daraufhin Vertreter des Stifts zusammen mit Veit von Drachsdorf zu einem Klärungsgespräch am 12 . Januar 1514 nach **Leipzig** geladen hat .

OPEN

ID bakfj:1290

Darin geht es um ein Treffen , das in **Leipzig** wegen eines Streits des Stifts mit Veit von Drachsdorf angesetzt ist .

OPEN

ID bakfj:1329

sachsen text

Sachsen text

Georg, Sachsen, Herzog

Herzog von Sachsen (albertinische Linie) seit 1500; entschiedener Gegner der Reformation




Figure 4: Entity-oriented search for the GND entity “4035206-7” (city of Leipzig), 

meet their specific identification or authorization requirements. *Shibboleth*<sup>13</sup> with its support for Single Sign-on (SSO) authentication in particular, has established itself for these purposes and is widely used, which makes it an ideal candidate for integration into the FCS. This implies to a greater extent that users have an academic background, but this is also reflected in access requirements for restricted resources, which is often *academic users only*. However, the distributed nature and different requirements for data protection and applications lead to very diversely configured identity providers, so available attributes to describe user identities are limited to email or anonymous identifiers<sup>14</sup> which only allows for relatively shallow verification.

The AAI specification extension proposes the integration of FCS clients (like the FCS Aggregator) as Service Providers where users can be authenticated if required. This authentication status will be forwarded securely and verifiably to FCS endpoints that require this information for access to their resources. The endpoint decides if the provided information meets the resource’s requirements, e.g. by checking for supported affiliations or user IDs, and return the results in case of success. In cases where this approach is not sufficient, dedicated diagnostics in the FCS protocol allow to state that results are present but can not be provided with an included reference to where users can request access. This “last resort” approach can at least provide some overview and an indication of potentially relevant data records at an institution for a user request.

The CLARIN FCS task force is currently reviewing a draft of the specification<sup>15</sup> and initial prototypes (including software libraries and deployment) are being evaluated to check if all requirements are met.

## 7 Supporting Resource-specific Fonts

The Unicode<sup>5</sup> standard comprises a wide variety of characters and symbols and enables the encoding of most texts. Nevertheless, as a global standard there are limitations on what is, can and will be included in the standard. In particular, historical and newer, non-standard applications make use of *Private Use Areas* in UTF-8 to map their own symbols into unused areas. This leads to issues for the general public when texts require custom fonts for correct rendering (e.g. font issues for lexical resources in Figure 5).

In the humanities, the necessity for custom fonts is particularly pronounced due to the diverse range of

<sup>13</sup>Identity management software, Shibboleth Consortium, <https://www.shibboleth.net/>

<sup>14</sup>The SAML attributes *eduPersonPrincipalName*, *eduPersonTargetedID*, or *mail*, were determined to be the minimal set that most Identity Providers can offer and whichever information is available will be used, see <https://tools.aai.dfn.de/entities/> for an overview in the DFN-AAI network.

<sup>15</sup>FCS AAI specification (draft), <https://clarin-eric.github.io/fcs-misc/fcs-aai-specs/fcs-aai.html>

scripts, symbols, and notational conventions used across different disciplines. Standardization processes, such as in the case of Unicode, often take considerable time, sometimes spanning several years. This delay poses challenges for projects that require immediate solutions for encoding and rendering specialized texts. Even when an encoding standard is eventually extended, integrating these updates into widely available fonts is not instantaneous, as font developers need additional time to implement and distribute the changes. Consequently, researchers and institutions must rely on custom fonts as interim solutions to ensure accessibility and readability of critical textual resources.

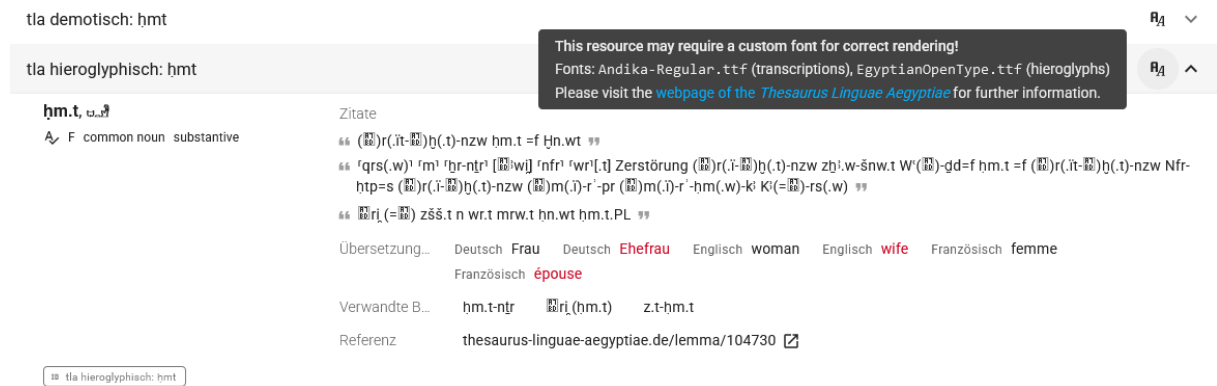


Figure 5: Font issues in *Thesaurus Linguae Aegyptiae* with information about custom fonts

A few interesting use-cases have already been identified where data providers plan to apply this new font extension: the *Thesaurus Linguae Aegyptiae* (BBAW, 2025), which requires fonts for both Egyptian texts and transcriptions, the *Hamburg Sign Language Notation System* (HamNoSys, Hanke, 2021), the *KompLett* font used for historical German texts, and *Landsmålsalfabetet* (Swedish Dialect Alphabet, Lundell, 1928) used for documenting dialects.

The proposal aims to provide a simple augmentation of the Endpoint Description to specify which resources may require certain fonts. Integrations in FCS clients can then either automatically apply the font in their front-end for results or offer a hint to users that further actions (e.g., downloading and installing a specific font) are required to enable correct rendering. Furthermore, the integration of information about required fonts and their sources not only enhances the representation and communication within FCS clients but also provides direct benefits for researchers who download FCS search results for further local analysis. Ensuring that font requirements are clearly documented allows scholars to work with the data in their own analytical environments without encountering unexpected rendering issues.

To ensure seamless integration of custom fonts within the FCS framework, a standardized approach to font metadata must be defined. This includes specifying font requirements within the metadata of the respective resources, enabling clients to recognize and apply the appropriate fonts dynamically. Additionally, font distribution mechanisms should be considered to facilitate user access, such as linking to web-hosted font repositories or embedding font files where permissible. Moreover, it is essential to address potential challenges related to the licensing and copyright constraints of custom fonts. Many specialized fonts are either proprietary or have restrictive licensing conditions that may prevent straightforward distribution. Therefore, FCS clients and providers must implement mechanisms to inform users about licensing terms and potential restrictions while offering alternatives, such as open-source fonts, whenever possible.

Currently, active work is underway to develop and refine this extension. Initial prototypes have already been built, demonstrating the feasibility of integrating font metadata into the FCS framework. Furthermore, discussions and exchanges with the community are ongoing to gather feedback and improve the approach based on real-world use cases and requirements. This will contribute to a more user-friendly and inclusive experience for diverse linguistic and historical text resources.

## 8 Improving Usability for all User Groups

An open, federated system like the FCS depends on a high degree of usability for all types of users, including developers, resource providers or end-users. In recent years, the FCS ecosystem has changed significantly to accommodate this goal. This includes a shift to popular working environments (GitHub/GitLab<sup>16</sup>) for supporting modern, open development processes, an expanded documentation based on easily editable formats (e.g. AsciiDoc<sup>17</sup>), provision of an increasing number of software libraries for a wide variety of use cases and the development of improved user interfaces.

To make the code base and documentation more accessible, it was moved to [GitHub.com](https://github.com). The working format for the FCS specification and other documents<sup>18</sup> was changed to AsciiDoc, to support improved editing, cross-linking and various output formats. Additional material like FCS Endpoint Development tutorials and extensive presentation slides<sup>19</sup> – focusing on different user groups – were created to ease development efforts.

Besides the documentation itself, various communication channels and activities are offered to support end-users, endpoint developers and operators. This includes the CLARIN Forum<sup>20</sup> to transparently publish news as well as allowing user interaction and feedback, workshops for FCS endpoint development (e.g., organised by the Saxon Academy of Sciences and Humanities in Leipzig (SAW) for Text+<sup>21</sup>) and support via help desks. More hackathons for developers and workshops for end-users focusing on specific usage scenarios are currently planned. All information material created in the process will be made available on the respective information channels as well.

On the software side, reference libraries that were initially developed in Java have now also been translated to Python to accommodate a greater variety of technology stacks. These libraries are continually being improved and extended. Numerous additional endpoint implementations in other languages exist as well – the open-source ones can be found in the aforementioned *Awesome FCS* list. To further assist FCS developers, various support tools are provided, many of them *dockerized* for an easier setup. A particularly important application is the **FCS Endpoint Validator**<sup>22</sup> which provides immediate feedback about the conformance of an endpoint to the FCS specification and which has been completely rewritten and extended to cover more test cases. It now offers additional features for configuration and reporting in response to evolving user requirements.

To facilitate end-user engagement, the FCS's central search application (the *FCS Aggregator*) has been revised several times since its initial release and has been continuously developed further. The latest major changes include an improved RESTful API<sup>23</sup> which can also be accessed from external applications. There is also a new Web component based on this API using the `Vue.js` framework. This component allows new forms of access to FCS endpoints and the presentation of results in an alternative web interface. In particular, this application supports a first integration of the EntityFCS based on the GND (cf. [Figure 4](#)), the improved presentation of lexical resources (cf. [Figure 2](#)), and also the option of integrating this new search interface into your own application with little effort. This is already implemented as part of the FCS integration in the central Text+ portal<sup>24</sup> (cf. [Figure 6](#)) and in the first repository websites that want to support search in their local resources<sup>25</sup> with a user-friendly interface.

## 9 Conclusion and Further Work

The Federated Content Search has proven and established itself as a platform that can react flexibly to changing or new requirements thanks to its open architecture and can therefore be easily integrated into

<sup>16</sup>For an overview of related repositories, see <https://github.com/clarin-eric/awesome-fcs>

<sup>17</sup>A plain text markup language for writing technical content, <https://asciidoc.org/>

<sup>18</sup>Overview page of compiled documents: <https://clarin-eric.github.io/fcs-misc/>

<sup>19</sup>Presentations slides for FCS endpoint development, <https://clarin.eu/fcsdevguide>

<sup>20</sup>All topics tagged with *fcs* in the CLARIN Forum, <https://forum.clarin.eu/tag/fcs>

<sup>21</sup>Blog post about the FCS Endpoint Development Hackathon, <https://textplus.hypotheses.org/9750> (in German)

<sup>22</sup>Official FCS Endpoint Validator for FCS / SRU protocol conformity and feature checks, <https://www.clarin.eu/fcsvalidator>

<sup>23</sup>OpenAPI-compliant description of the FCS Aggregator REST API, <https://contentsearch.clarin.eu/openapi.json>

<sup>24</sup>FCS Web Component integration in Text+ webpage, <https://text-plus.org/en#action-open-search?tab=content>

<sup>25</sup>FCS integration at the SAW Leipzig repository with a filtered resources list, <https://repo.data.saw-leipzig.de/en#open-fcs>

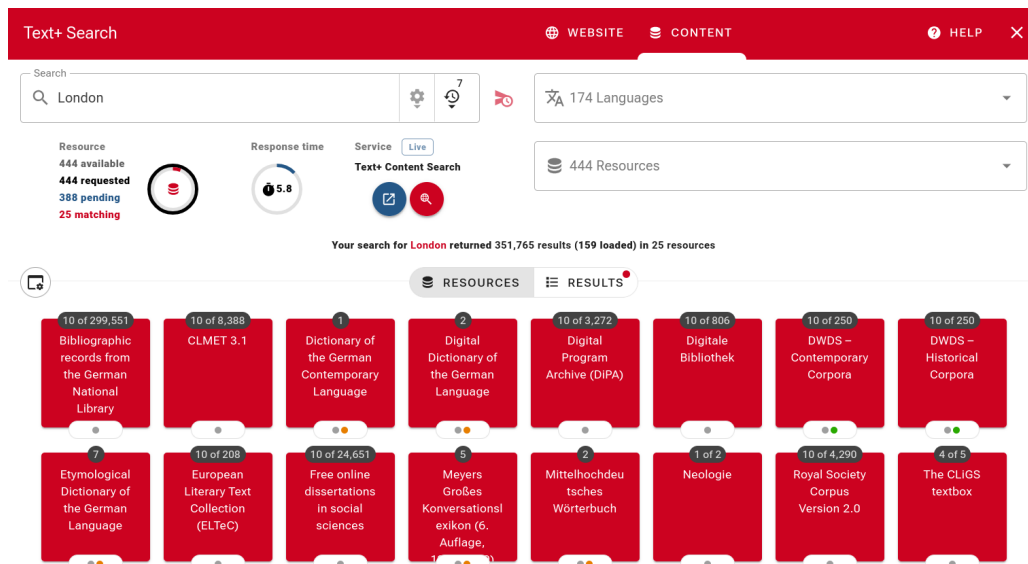


Figure 6: Result overview for a FCS query on the webpage of the Text+ project, [🔗](#)

new research contexts. In recent years, specific application-driven usage scenarios have led to various lightweight extensions and customisations, many of which are already in active use. The work presented not only illustrates the dynamics of the development to date, but also forms the basis for further short and medium-term planning. The focus for the coming years will be on further lowering the entry barrier for users and developers and adapting current developments in the area of large language models, such as the use of the FCS in the context of Retrieval Augmented Generation (RAG).

More FCS extensions are currently being worked on, but are still in the user requirements analysis phase or consist of only initial prototypes for evaluation and testing purposes. The potential support for queries on syntactic structures (e.g., in dependency treebanks) is a fairly recent development. A proposal for such a new type of query language might lead to a new *SyntacticFCS* extensions. Another aspect under investigation is an improved description of results through structured metadata. The FCS specification is currently very limited on how much additional metadata can be provided and used by software clients and end users. Data providers with resources that, for example, aggregate texts from various sources such as newspaper collections, require ways to enrich individual results with more descriptive metadata, including publication and release date, title, or author information.

In any case, for each extension, a decision must be made as to whether the specific need justifies the corresponding work, for example whether a sufficient number of resource providers can or want to support the respective type of information or the respective use case. To ensure the functionality and interoperability of the entire FCS infrastructure, issues such as backward compatibility and mechanisms for explicitly opting out of individual functionality play an important role here.

## Acknowledgments

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* and the *CLARIN FCS taskforce* for their continuous work.

## References

- Ayoola, T., Fisher, J., & Pierleoni, A. (2022, July). Improving entity disambiguation by reasoning over a knowledge base. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2899–2912). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.210>
- Balog, K. (2018). *Entity-Oriented Search* (Vol. 39). Springer Cham. <https://doi.org/10.1007/978-3-319-93935-3>
- BBAW. (2025). Thesaurus Linguae Aegyptiae [Accessed: 2025-04-09]. <https://thesaurus-linguae-aegyptiae.de>
- Eckart, T., Herold, A., Körner, E., & Wiegand, F. (2023). A Federated Search and Retrieval Platform for Lexical Resources in Text+ and CLARIN. *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, 280–292. [https://elex.link/elex2023/wp-content/uploads/elex2023\\_proceedings.pdf](https://elex.link/elex2023/wp-content/uploads/elex2023_proceedings.pdf)
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books. <https://mitpress.mit.edu/9780262561167/>
- German National Library (DNB). (2024). The Integrated Authority File (GND) [Accessed: 2024-06-24]. [https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd\\_node.html](https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html)
- Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with wikipedia [Artificial Intelligence, Wikipedia and Semi-Structured Resources]. *Artificial Intelligence*, 194, 130–150. <https://doi.org/https://doi.org/10.1016/j.artint.2012.04.005>
- Hamp, B., & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. <https://aclanthology.org/W97-0802>
- Hanke, T. (2021). Hamnosys. <https://doi.org/10.25592/uhhfdm.9725>
- Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. (2009). ISOcat: Remodelling metadata for language resources. *IJMSO*, 4, 261–276. <https://doi.org/10.1504/IJMSO.2009.029230>
- Körner, E., Eckart, T., Kretschmer, U., Herold, A., Wiegand, F., Michaelis, F., Bremm, M., Cotgrove, L., Trippel, T., Rau, F., Klee, A., Werning, D., Blöse, D., & Zinn, C. (2024). *Federated Content Search for Lexical Resources (LexFCS): Specification*. <https://doi.org/10.5281/zenodo.7849753>
- Lundell, J. A. (1928). The swedish dialect alphabet. *Studia Neophilologica*, 1(1), 1–17. <https://doi.org/10.1080/00393272808586721>
- OASIS. (2013). *searchRetrieve v1.0*. Organization for the Advancement of Structured Information Standards. <http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/searchRetrieve-v1.0-part0-overview.html>
- Qi, L., Yongyi, H., Defu, L., Zhi, Z., Tong, X., Che, L., & Enhong, C. (2024). Unimel: A unified framework for multimodal entity linking with large language models. <https://arxiv.org/abs/2407.16160>
- Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmänn, M., & Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*, 5. <https://doi.org/10.17175/2020.006>
- 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*. <https://www.clarin.eu/content/federated-content-search-core-specification>
- Schwarz, P., & Barth, F. (2024, March). Classification and linking of named entities. <https://doi.org/10.5281/zenodo.10893761> Workshop contribution in Pollin, Christopher, et al. "Workshop generative KI, LLMs und GPT bei digitalen Editionen".
- Stehouwer, H., Durco, M., Auer, E., & Broeder, D. (2012). Federated Search: Towards a Common Search Infrastructure. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3255–3259. <http://www.lrec-conf.org/proceedings/lrec2012/pdf/524.Paper.pdf>

- van Uytvanck, D., Olsson, L.-J., Schonefeld, O., Eckart, T., Körner, E., Kisler, T., Fischer, P. M., & Illig, E. M. (2017). *CLARIN Federated Content Search (CLARIN-FCS) – Core 2.0*. <https://office.clarin.eu/v/CE-2017-1046-FCS-Specification-v20230426.pdf>
- Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10), 78–85. <https://doi.org/10.1145/2629489>