# Federated Search: Towards a Common Search Infrastructure

**Herman Stehouwer[†], Matej Durco[‡], Eric Auer[§], Daan Broeder[¶]**

[†,§,¶] Max Planck Institute for Psycholinguistics, [‡] ICLTT
[†,§,¶] Nijmegen, The Netherlands, [‡] Vienna, Austria
[†,§,¶]{herman.stehouwer, eric.auer, daan.broeder}@mpi.nl, [‡] matej.durco@assoc.oeaw.ac.at

**Abstract**

Within scientific institutes there exist many language resources. These resources are often quite specialized and relatively unknown. The current infrastructural initiatives try to tackle this issue by collecting metadata about the resources and establishing centers with stable repositories to ensure the availability of the resources. It would be beneficial if the researcher could, by means of a simple query, determine which resources and which centers contain information useful to his or her research, or even work on a set of distributed resources as a virtual corpus. In this article we propose an architecture for a distributed search environment allowing researchers to perform searches in a set of distributed language resources.

**Keywords:** search, infrastructure, distributed

## 1. Introduction

In the recent years the focus in the field of language resources has shifted from the creation of resources to improving the accessibility of existing resources. Accessibility implies work on harmonizing the data formats and access methods. Harvesting and searching in metadata is a basic principle of many infrastructure projects (e.g., CLARIN, DARIAH, METANET). However searching in the content of the resources is not a solved problem. Besides the legal issues (which are beyond the scope of this article), there are technical issues. If the data is accessible from the web, it is usually only accessible using a custom search system. Queries are generally not compatible between different search systems. Furthermore, the results returned by the different systems vary greatly in their formatting (Johnson, 2002; Skiba, 2009; Wittenburg et al., 2010). In order to tackle this problem we propose a federated search infrastructure. This infrastructure is developed within the CLARIN[1] infrastructure (Váradi et al., 2008) . Using this infrastructure the researcher will be able to quickly see which corpora, or resources contain potentially useful information by performing a content-search. That is, the distributed search is meant to search within the content of the resources (sentences, transcriptions, glossings, gesture annotations, etc.).

The goal of the federated search is not to replace the specialized search engines, but to give a quick, global, overview of fitting resources. The presence of such a search engine enables serendipity. The federated search architecture also allows the researcher to search specific data-sets or parts of corpora. We note that the search engines specific to the resources themselves often offer extra, resource specific, search options.

This article is structured as follows. In Section 2. we will briefly cover the initial state of some of the participating repositories. In Section 3. we describe the search in frastructure that we are creating using these protocols and how it can be used by the researcher. In Section 4. we will cover the SRU/CQL specification and briefly compare it with alternative protocols, before we elaborate on the extension thereof to fit our purpose. Finally, in Section 5. we describe the next steps in developing this infrastructure and how we hope to move towards a combined European search infrastructure.

## 2. Initial Situation

In this section we describe the five archives (with their corresponding search systems) that currently participate in the prototypical search federation. Each of these archives provide access to searching one or more corpora using SRU/CQL. This way it quickly becomes clear for instance which corpora contain resources with specific content, such as 'cow'.

First, (Stehouwer and Auer, 2011; Wittenburg et al., 2010) give a recent description of the state of The Language Archive (TLA) and the available search methods. TLA contains mainly currently spoken linguistic data, oftentimes accompanied by video, hosting data from many linguistic preservation projects, linguistic studies and psycholinguistic experiments. Overall the archive contains circa $160,000$ annotation files for more than $200,000$ audio or video recordings. TLA offers a variety of methods to browse, search, and leverage the large body of data.

Within the Federated Search infrastructure the TLA offers five corpora for search: (1) CGN, or the Corpus Gesproken Nederlands (Corpus of Spoken Dutch), (2) the ESF corpus, a corpus of second language learning, (3) the IFA corpus, a corpus of hand-segmented Dutch speech, (4) the Childes corpus, a corpus of children speech within first language acquisition, and (5) the Talkbank corpus.

Second, C4 (Dittmann et al., in prep) is a collaborative project by four partners (DWDS Berlin, AAC Vienna, CHTK Basel, and EURAC Bozen) to construct a distributed corpus of standard varieties of modern German. It can be seen as direct predecessor of the current federated search effort, as it is a distributed implementation. Each partner offers its data on their own server, and the query of the user is distributed to the four servers and the results are com-

[1] See *www.clarin.eu* on the web.

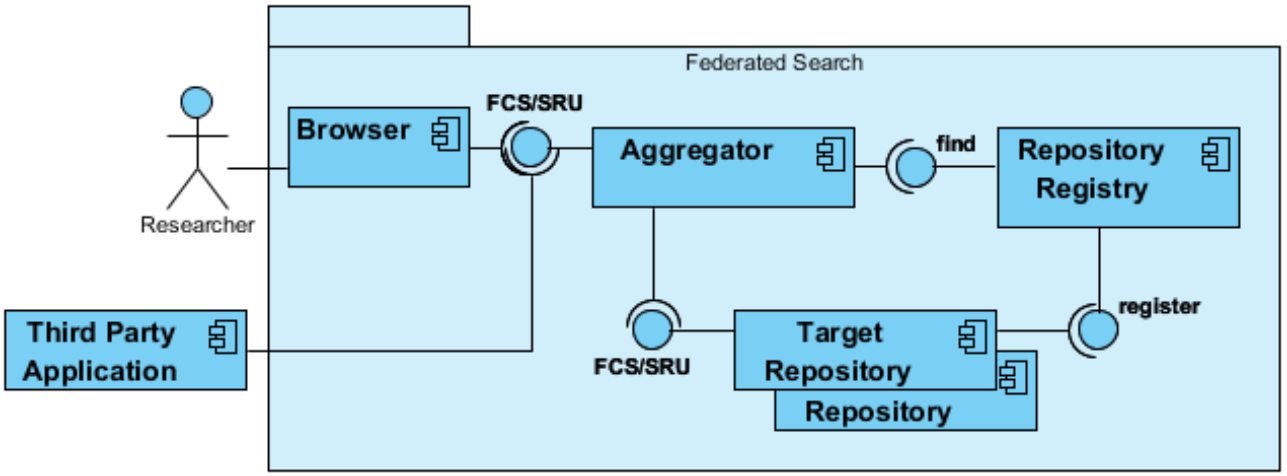


Figure 1: A diagram showing the main components of the architecture.

bined before being presented to the user. The limitation of C4 is, that the distributed aspect is realised by the search engine ddc-concordance (Sokirko, 2003). That is, this engine is installed by every party and requires the usage of a certain inner structure of the corpora. While this was feasible in the limited scope of the project, it is an unacceptable constraint when dealing with a large heterogeneous resources landscape.

Third, the Mimore tool enables researchers to investigate morphosyntactic variation in Dutch using three databases; (1) DynaSAND (Barbiers et al., 2006) the *dynamic syntactic atlas of the Dutch dialects*, containing results from oral fieldwork on Dutch dialects in some 300 locations spread throughout the Netherlands, Belgium, and a small part of northern France, (2) DiDDD (Corver et al., 2005) *Diversity in Dutch DP Design*, containing results obtained using similar methodology to DynaSAND from some 200 locations pertaining morphosyntactic variation in nominal groups, (3) GTRP (Goeman and Taeldeman, 1996) the *Goeman, Taeldeman, van Reenen Project*, containing data from some 600 locations from the Dutch language area.

Fourth, the INL makes available a search option on the Gysseling corpus. The Gysseling corpus consists of thirteenth century texts that where used as the basis for the dictionary of early Dutch Toponymy (Gysseling, 1960)

Fifth, DANS makes available the Lieffering corpus. The Lieffering corpus consists of old Dutch works collected for (Lieffering, 2007).

## 3. Architecture

The architecture used for the federated search service is quite straightforward. It consists of four major components, which we will list here in bottom-up order.

The first component is the *target repository*, a system hosting a set of resources, making them available for search via an agreed upon protocol (discussed in Section 4.).

The second component is the *aggregator*, which is responsible for two things: 1) distributing the user query to the appropriate target repositories, and 2) combining the responses from those target repositories into a merged result-set. The aggregator should provide a protocol conformant interface as well. Unified interfaces will allow the clients to easily switch between accessing a single repository and accessing the aggregator.

The third component is the *browser* or the user interface. This component offers the user an attractive and easy-to-use web interface for making queries to (a selection of) target repositories. This component is also responsible for displaying.

The fourth component is the *repository registry*, which is responsible for keeping track of all the available target repositories.

## 4. Protocol

As we mentioned in the previous section the target repositories implement a common protocol to provide search access to their data. For the common protocol we decided to use a modified, but interoperable, version of the SRU/CQL protocol.

SRU/CQL is the communication protocol and query language proposed by the Library of Congress. It is a simplified, XML- and HTTP-based successor to Z39.50 (Lynch, 1991), which is very widely spread in the library networks. Libraries were the early adopters and driving force in the field of search federation even before the era of internet, starting collaborative efforts in mid 70s (Linked Systems Project (Fenly and Wiggins, 1988)).

SRU/CQL was introduced 2002 (Morgan, 2004). Coming from the libraries world, the protocol has a certain bias in favor of bibliographic metadata. However, the protocol is defined in a very generic way, with a strong focus on extensibility. It is equally suitable for content search. We refer to (Morgan, 2004) for a precise definition of the protocol.

The protocol part (SRU) defines three major operations: 1) *explain*: in which the target repository announces its particular configuration (e.g. available indices), 2) *scan*: informing about terms available in/for given index, and 3) *searchRetrieve*: returning a search result based on a CQL query.

The query language part (CQL - Context Query Language) defines a relatively complex and complete query language. The decisive feature of the query language is its inherent

extensibility allowing to define own indexes and operators within the so called 'context sets'. We provide some telling examples:
Simple term query:

*wolf*

phrase:

*"Who's afraid of"*

index-search:

*dc.title any "open access"*

index + boolean search:

*dc.date > 1900 and dc.date < 1910*

**Conformance Levels**

It is clear that no target repository will be able to provide a full protocol conformance right away. However, a partial implementation can already be useful. SRU proposes Conformance levels 0 to 2. While these conformance levels can serve as a guide, a more fine-grained assessment of the features is advisable. We propose to break the functionality down to individual features that can be implemented and tested separately. The basic level (conformance level 0) is to support a simple keyword search. This is straightforward to implement on top of an existing search engine, enabling centers to quickly and easily join the search federation with the basic functionality.

**Extensions to the protocol**

In the first step the implementing parties concentrate on setting up the connectivity with a simple keyword search. However, individual target repositories have the possibility to implement more advanced queries. CQL provides great means for quite advanced queries. Based on the notion of context sets it allows to introduce new indices, relations and modifiers. However, CQL foresees the indices to be defined as a static list. We need to evaluate, based on use if this is sufficient or if the proposed system needs a more flexible way of announcing the indices.
The most straight-forward solution is to introduce a new context set for Federated Content Search (FCS) and add indices requested by individual target repositories. However it is important that the system is not flooded with a large number if disjunct indices. It is up to the target repositories to define their indices in an agreed upon manner. This is where the data category registry *ISOcat* (Kemps-Snijders et al., 2009) is the obvious choice as a stable common ground. It stores definitions of concepts that can be used as search indexes and will be defined as a separate context set - isocat. Consequently, the isocat-indexes are to be preferred and fcs-indexes should only be used, if there is no isocat equivalent. Using these two context sets, if will be possible to formulate queries like:

*(fcs.word = cow)*
*isocat.word = cow*
*isocat.lemma = fly*
*isocat.partOfSpeech = noun*
*fcs.pos regexp N.**
*isocat.lemma = fly AND fcs.pos = noun*

Another important aspect within a federated search is the need to restrict the search to a part of the available material. For this purpose we propose to introduce an extension parameter, that allows the client to indicate which parts of the corpus or parts of the dataset he or she wants to query.
Finally, we propose a generic schema for structuring the results. While SRU defines an envelope for the result format, it allows for any data inside the individual result (inside the *<sru:recordData>* element). The proposed schema shall provide a generic frame to fit in the various types of data that we expect, providing a means to distinguish between metadata, content and various data-views, i.e. types of content. Specifically it allows: a) any metadata about the matched record, including reference to a CMDI metadata-record (Broeder et al., 2011), b) separate (metadata) description of a full dataset or corpus and a part of that dataset or that corpus, c) various views on the results of the query (such as keyword in context), and d) to provide links to any of the above (i.e. reference the data instead of including it directly in the result).
As mentioned, the results each contain dataviews within the SRU-defined *<sru:recordData>* element. We allow the presence of multiple dataviews, representing the same part of the data on which the hit was found. One of these dataviews must be in the keyword in context format. Furthermore we support the use of dataviews for representing metadata, geographic information, full-text without annotation and full-text with annotation.

**Protocol of the Aggregator**

The aggregator should itself provide a SRU-conforming search API. Unified interfaces will allow the implementing client services to switch between accessing a single repository and the aggregator. Furthermore a REST-style handling of requests allows the Aggregator to easily cache the results and it allows the user to distribute and share results by simplyf sending a URL.
However, in a distributed environment, when dealing with potentially slow or non-responding targets, a stateless atomic request-response can pose a problem. Although this can be solved with time-outs, there is a trade-off between getting the most results and letting the user wait. Thus the aggregator should provide an additional session-based interface targeted at browsers. That would allow to return intermediate partial results without delay, which would be a substantial quality improvement in user experience. The user can already inspect the partial results, while waiting for the full response.

## 5. Status and Future

In this article we presented an architecture for distributed search environment. To reiterate, we have defined a modification to SRU/CQL that makes it suitable for a federated content search. We stress that producing a base-level implementation of the protocol and query language is low in the amount of effort required.
The archives described in Section 2. and also a few other CLARIN content providers already implement a basic FCS/SRU-endpoint. There is also a first version of the

Repository Registry, storing and publishing the informations about individual endpoints. More work needs to be done on harmonizing the result formats and the user interface / browser for accessing the search functionality.

In general producing a federated search offers some significant advantages over a central search index. These include the possibility of incorporating very diverse sets of resources without having to store them in a central database as well as the ease of scaling to a large number of endpoints without reducing overall performance too much.

In the future we will proceed with implementing these ideas further. First of all we will endeavor to make available bindings to popular index systems, such as Lucene and invite data providers to join the initiative. Second, the system will be extended to integrate more with the rest of the CLARIN infrastructure such as the existing metadata search, and the virtual language observatory (Uytvanck et al., 2010). Thirdly, integration with access rights systems will have to be provided, at the moment only publicly searchable corpora are made available. To implement this we need to solve issues related to delegating the users identity to the search engines involved. Finally, we plan to use ISOcat to mediate between the intention of the user and the representation used within the different corpora.

## 6. References


E. Auer, P. Wittenburg, H. Sloetjes, O. Schreer, S. Masneri, D. Schneider, and S. Tschöpel. 2010. Automatic annotation of media field recordings. In C. Sporleder and K. Zervanou, editors, *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, pages 31–34, Lisbon. University de Lisbon.

S. Barbiers, H. Bennis, G. De Vogelaer, M. Devos, and M. van der Ham. 2006. Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND).

Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to xml interoperability the component metadata infrastructure (cmdi). In *Proceedings of Balisage: The Markup Conference. Balisage Series on Markup Technologies*, volume 7, Montréal, Canada.

N. Corver, M. van Koppen, H. Kranendonk, and M. Rigterin. 2005. The Noun Phrase: Diversity in Dutch DP Design (DiDDD). In *Scandinavian Dialect Syntax*, pages 73–85, Tromso, Norway.

Henrik Dittmann, Matej Durco, Alexander Geyken, Tobias Roth, and Kai Zimmer. in prep. Korpus C4 – a distributed corpus of German varieties. In Thomas Schmidt and Kai Wrner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, number 14 in Hamburg Studies on Multilingualism. Amsterdam: Benjamins.

J. G. Fenly and B. Wiggins. 1988. *The Linked Systems Project: a networking tool for libraries*. OCLC (Online Computer Library Center), Columbus, OH, USA. Fenly:1988:LSP:60988.

A. Goeman and J. Taeldeman. 1996. Fonologie en morfologie van de Nederlandse dialecten. In *taal en tongval 48*, pages 38–59, Gent, Belgium.

M. Gysseling. 1960. *Toponymisch woordenboek van België, Nederland, Luxemburg, Noord-Frankrijk en West-Duitsland (vóór 1226)*. Belgisch interuniversitair centrum voor neerlandistiek.

H. Johnson. 2002. The Archive of the Indigenous Languages of Latin America: Goals and Visions. In *Proceedings of the Language Resources and Engineering Conference*, Las Palmas, Spain.

M. Kemps-Snijders, M. Windhouwer, and P. Wittenburg. 2009. ISOcat: Remodeling metadata for language resources. In *International Journal of Metadata, Semantics and Ontologies (IJMSO)*, volume 4 (4), pages 261–276.

M. Kemps-Snijders, T. Koller, H. Sloetjes, and H. Verweij. 2010. LAT bridge: Bridging tools for annotation and exploration of rich linguistic data. In N. Calzolari, B. Maegaard, J. Mariani, J. Odijk, K. Choukri, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 2648–2651. European Language Resources Association (ELRA).

A. Lieffering. 2007. *The French comedy 1749-1793*. Koninklijke Vereniging voor Nederlandse Muziekgeschiedenis.

Clifford A. Lynch. 1991. The Z39.50 information retrieval protocol: an overview and status report. *SIGCOMM Comput. Commun. Rev.*, 21(1).

E.L. Morgan. 2004. An introduction to the Search/Retrieve URL Service (SRU). *Ariadne*, July.

R. J. F. Ordelman, W. F. L. Heeren, F. M. G. de Jong, M. A. H. Huijbregts, and D. Hiemstra. 2009. Towards Affordable Disclosure of Spoken Heritage Archives. *Journal of Digital Information*, 10(6):17, December.

J. Ringersma, C. Zinn, and A. Koenig. 2010. Eureka! User friendly access to the MPI linguistic data archive. In *SDV - Sprache und Datenverarbeitung/International Journal for Language Data Processing*.

R. Skiba. 2009. Korpora in der Zweitspracherwerbsforschung: Internetzugang zu Daten des ungesteuerten Zweitspracherwerbs. In B. Ahrenholz, U. Bredel, W. Klein, M. Rost-Roth, and R. Skiba, editors, *Empirische Forschung und Theoriebildung: Beiträge aus Soziolinguistik, Gesprochene-Sprache- und Zweitspracherwerbsforschung: Festschrift für Norbert Dittmar*, pages 21–30.

A. Sokirko. 2003. DDC: A search engine for linguistically annotated corpora. In *Proceedings of Dialogue*, Protvino, Russia.

J.H. Stehouwer and E. Auer. 2011. Unlocking Language Archives Using Search. In *Language Technologies for Digital Humanities and Cultural Heritage*.

P. Trilsbeek, D. Broeder, T. Valkenhoef, and P. Wittenburg. 2008. A grid of regional language archives. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-

conf.org/proceedings/lrec2008/.

D. Van Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, and M. Gardelleni. 2010. Virtual language observatory: The portal to the language resources and technology universe. In N. Calzolari, B. Maegaard, J. Mariani, J. Odijk, K. Choukri, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 900–903. European Language Resources Association (ELRA).

Tamás Váradi, Peter Wittenburg, Steven Krauwer, Martin Wynne, and Kimmo Koskenniemi. 2008. Clarin: Common language resources and technology infrastructure.

P. Wittenburg and P. Trilsbeek. 2010. Digital Archiving – a necessity in documentary linguistics. In G. Senft, editor, *Endangered Austronesian and Australian Aboriginal languages: Essays on language documentation, archiving and revitalization*, pages 111–136. Canberra: Pacific Linguistics.

P. Wittenburg, P. Trilsbeek, and P. Lenkiewicz. 2010. Large multimedia archive for world languages. In *Proceedings of the 2010 ACM Workshop on Searching Spontaneous Conversational Speech, Co-located with ACM Multimedia 2010*, pages 53–56. Association for Computing Machinery, Inc. (ACM).