Title	Workshop on Federated Content Search, Copenhagen, 2013-04-24
Version	1
Author(s)	Thorsten Trippel
Date	2013-05-02
Status	Draft
Distribution	Participants
ID	CE-2013-0146

Workshop on Federated Content Search, Copenhagen, 2013-04-24

1 Participants, time, place

Participants: Ronald Haentjens Dekker, Christoph Draxler, Matej Ďurčo, Thomas Eckart, Guido Gerritsen, Tobias Gradl, Hanna Hedeland, Ivajlo Ivanov, Bart Jongejan, Neeme Kahusk, Wim Kok, Tomas Krilavicius, Karlheinz Moerth, Jan Niestadt, Tonis Nurk, Jan Odijk, Lene Offersgaard, Yana Panchenko, Maciej Piasecki, Oliver Schonefeld, Mitchell Seaton, Kiril Simov, Ivan Skoglund, Pavel Stranak, Thorsten Trippel, Dieter Van Uytvanck, Marta Villegas, Freddy Wetjen

CLARIN-ERIC office: Thorsten Trippel

Chair: Christoph Draxler

Date and Time: 2013-04-24, 09:30 – 16:30 CET

Location: University of Copenhagen

2 Agenda

1. Welcome by the chair of the Standing Committee of Clarin Technical Centers Dieter van Uytvanck
2. Roll call
3. Overview on Federated Content Search
4. Clarin Federated Content Search: Specification and Software Components (Oliver Schonefeld)
5. BlackLab: a researcher's best friend (Jan Niestadt)
6. FCS at UPF (Marta Villegas)
7. CLARIN-D Federated Content Search Aggregator (Yana Panchenko)
8. FCS@CLARIN-AT (Matej Ďurčo)
9. DARIAH general search (Tobias Gradl)
10. Example queries for Federated search (Jan Odijk)
11. Summary (Christoph Draxler)

3 Welcome by the chair of the SCCTC

The chair of the Standing Committee of Clarin Technical Centers, Dieter van Uytvanck, welcomed the participants and explained the motivation for the workshop. The workshop is to provide an overview of the current state of the specification of federated content search within CLARIN and the involved protocols and to locate the gaps. Additionally it should show some experiments with FCS and provide an opportunity to discuss use cases for possible searches.

4 Roll call

During the roll call all participants introduced themselves and their affiliation.

5 Overview on Federated Content Search

Chrisoph Draxler gave an overview of federated content search by showing the CLARIN-D implemented FCS aggregator and searching over various centers. He showed that a full text search is possible, but that some queries are not yet satisfying because for example the text and the markup are not sufficiently distinguished. For finding material containing specific forms this current implementation seems very useful.

6 Clarin Federated Content Search: Specification and Software Components

Oliver Schonefeld presented the overview and architecture of the components and the library for Federated Content Search implemented by him. This Federated Content Search uses SRU (Search/Retrieve via URL) with some proposed extensions for the use within the CLARIN FCS. Different views are defined in the FCS for example KWIC and geolocation, more views can be added.

The current specification document available via the CLARIN TRAC system requires background knowledge in SRU/CQL (in FCS: Contextual Query Language). However, suggestions for the improvements of the specification document are welcomed.

At present, no aggregation of search results is available, that means cross-endpoint answers not possible. To test endpoints, a service is available at http://clarin.ids-mannheim.de/srutest (requires authorization). This test will be performed prior to inserting the SRU endpoint into the center registry.

The discussion after the presentation showed that FCS will not replace specialized search engines because the specific search engines are working for specific search questions and an integration into one system would result in an unmaintainable query language. For the present purposes it seems most relevant to show use cases that are feasible with the current data and close to today's implementation to raise interest in FCS.

The slides of this presentation are available in PDF.

7 BlackLab: a researcher's best friend

Jan Niestadt reports on BlackLab, a corpus query engine base on Apache Lucene. The search supports various input formats, more input formats may become available upon request. In the future, BlackLab may include the corpus query processor of the Sketch Engine. The native query language is based on CQL (Corpus Query Language), which is mapped onto the Lucene query syntax.

The discussion showed that the crucial problem for such applications to be integrated in FCS is the translation of SRU/CQL (Contextual Query Language) into the specialized query languages, here CQL (Corpus Query Language) used by the backend without losing functionality of the specialized query languages. It seems possible that SRU/CQL (Contextual..) is not rich enough to express everything from CQL (Corpus...) in a query. This would limit the options for FCS.

8 FCS at UPF

Marta Villegas showed the FCS at UPF. Beyond the slides that are made available, she addressed the central problem of the possible massive amount of data resulting from a search operation which requires an additional statistical analysis to manage the results. This will require further work.

9 CLARIN-D Federated Content Search Aggregator

In her presentation, Yana Panchenko presented an FCS aggregator that serves as the user frontend to the FCS, i.e., the aggregator addresses the SRU/CQL endpoints in a search operation and returns the results to the user. The discussion showed that simple searches with the aggregator can be used for quality assurance. At present the aggregator as such is not planned to serve as a SRU/CQL endpoint.

10 FCS@CLARIN-AT

In his presentation, Matej Ďurčo demonstrated how metadata and content search can be combined. This includes a reference to SRU/SQL endpoints in CMDI data to query the specific resource.

11 FCS in DARIAH

Tobias Gradl demonstrated the DARIAH generic search approach. According to the discussion, uniform access seems not feasible but either subject specific deep or flat queries are used.

He demonstrated the search at http://demo2.dariah.eu:8080/search/search/extended .

Though the data is available in many indexes, the prototype only harvests Dublin Core.

12 Example queries for Federated search

To discuss the options for FCS Jan Odijk presented use cases that would be of interest to linguists using FCS. It turned out that one problem is the missing semantic interoperability of for example tags that are not mapped onto a data category registry.

Very helpful for users would be a Query-by-example approach in which users would see executed queries and could use them to create their own queries.

The discussion showed that the creation of a user interface pose a special problem, as the users do not have been exposed to a prototypical interface which makes it easy to create complex queries. Interfaces with buttons could be relevant in domain specific interfaces, but these often restrict the power of the query language. A graphical interface that shares features of a mindmap would be one option. It would also be necessary to distinguish different kinds of search interfaces, for example lexicon, treebanks, corpora, etc. if the resource specific structures should be useable via FCS.

Pavel Straňák added to the discussion that the repository may not be the best place for accessing the data but that specialized clients could be better suitable. This would also allow to incorporate licenses. He also pointed out that a too great variability without a semantic mapping could render a search useless.

During the discussion an additional function of FCS was stressed, namely reaching out to the user community. Only in a separate step it should be considered to integrate further options. Taking the experience from Prague, only few users use the structures for deep treebank-based searches, and integrating this into a FCS may take too much of an effort to a user community that can already use existing complex query tools.

13 Summary of the discussion

At the end Christoph Draxler summarized the topics discussed during this workshops using the following individual items:

1. "Metadata police": it seems to be necessary to get consistency into CMDI and ISOcat by harmonization. This also includes to be more strict with obligatory fields and closed vocabularies. Profiles that are available should

be propagated, and tools that make it easy to use need to be available. This also may include scripts to convert Excel-tables to CMDI

2. Metadata vs. Data: It seems that in the context of FCS the lines between metadata and content are blurred.

3. "Users": For the further development of FCS the user needs to be specified, i.e., if the data is used by humans or machines. Use cases should be used to demonstrate what is possible with FCS and show the limits.

4. Data views: A search operation alone may not be sufficient but, data views can be seen as a resource, which would allow multiple viewers.

5. Large results: The mass of possibly returned data

6. Query formulation: for FCS the expectations regarding expressivity of queries should not be too high, but full text searches should be developed that show some useful output; existing software that is FCS-enabled like BlackLab could be propagated further.

7. Supporting output formats: For further processing of results, simple formats such as CSV should be made available.

The Standing Committee of CLARIN Technical centers will form a small taskforce looking at the documents for CLARIN. The goal is to provide for a:

1. setup for infrastructures at individual centers, this should allow a basic integration with the aggregator.

2. continuous work on standards and classification.