TEXT

CLARIN Federated Content Search 2.0 – state of the union

Leif-Jöran Olsson

Språkbanken Text (SBX), University of Gothenburg, CLARIN ERIC,
Swe-Clarin

2020-03-31

This presentation covers

► The uptake of the new major release of the CLARIN federated content search, CLARIN-FCS 2.0, that brought groundbreaking new features and capabilities to the linguistic search of corpora.

► The new proposal for accessing restricted resources.

► The upgrade/replacement of internal core in the dropwizard framework

► UI changes

► Threats

Uptake

- ▶ The CLARIN-FCS version 2.0 final specification has been out for three years
- ▶ The Aggregator has been supporting FCS v2.0 for almost 4 years
- ▶ In the last year we have gone from 77 to around 90 collections
- ▶ (cf roughly 1150 to 1235 collections for FCS 1 or legacy)

TEXT Introductory pitch

The CLARIN Federated Content Search (CLARIN-FCS) introduces an *interface specification* that decouples the *search engine* functionality from its *exploitation*, i.e. user-interfaces, third-party applications, to allow services to access heterogeneous search engines in a uniform way.

Augment not replace

This means the CLARIN-FCS is augmenting your search engine – not replacing it. And by still keeping a low threshold version 2.0 introduces a few important new features:

► A new Query Language

► A matching display of query results, the AdvancedDataView, ADV, with layer capabilities

► Backwards compatibility to earlier versions

These new additions to the CLARIN-FCS will not only enhance the power user experience and possibilities when performing queries from repositories, but also that less experienced users will find it easier to explore different corpora.

► Authentication via JWT (JSON web token)

► Adding dedicated element to
EndpointDescription

► Adding diagnostics

► Adding tests

We will hear more from Gregor Middell about the
BBAW attempt on doing the authentication

Suggested EndpointDescription addition (updated)

```
<Resources>
    <Resource>
        ...
        <Languages>
            <Language>swe</Language>
            <Language>deu</Language>
        </Languages>
        <AvailabilityRestriction>
            <RestrictionRequirement>requirementName</RestrictionRequirement>
        ...
        </AvailabilityRestriction>
        ...
    </Resource>
</Resources>
```

Figure: Suggested AvailabilityRestriction addition (in red) for CLARIN-FCS 2.0

TEXT

Suggested EndpointDescription addition (description, updated)

► The text node *requirementName* in *RestrictionRequirement* is either of {authOnly, personalIdentifier}.

► PersonalIndentifier for Shibboleth comes from a mapping of eduPersonPrincipalName or eduPersonTargetedID or email to "userID". "userID" is then passed to the endpoint.

► If you define the *AvailabilityRestriction* in the top-level *Resource* element the restriction will be valid for the complete tree hierarchy of *Resource* elements.

► From the backwards compatibility perspective this means that if you do not define the *AvailabilityRestriction* element all resources will be available for searching through the endpoint as expected.

Upgrading core of dropwizard

► Has been an ongoing chase of replacing dependencies for dropwizard 1.0

► Going to dropwizard 2 requires extensive rework of internals (Willem finally took on the task)

► Still need more tests

▶ Adding autocompletion where possible in textbased FCS-QL

▶ Adding authentication for restricted resources

▶ Making some things more smooth

TEXT Other changes

▶ Extending grammar with changes for the syntactic search capability

▶ Aggregation examples

▶ Upgrading further libraries

Going forward and still keeping backwards compatibility

► Even with the new features of version 2.0 the threshold to participate is still low. The backwards compatibility gives you as a Centre a smooth transtion to the new features and capabilities at your own convenience.

► Addition of access to restricted resources (very few plan to do it, but uptake might increase)

► Slow uptake

Summary

Altogether, we see CLARIN-FCS 2.0 as an easy to use, easy to maintain, federated search that will make the language resource exploration more open to everyone, no matter if you are a power or less experienced user, or a resource provider.

Version 3.0 and beyond

With this strong foundation of The Aggregator and ADV laid in CLARIN-FCS 2.0, we can now develop the ways to formulate and present queries and their results, making a good foundation for addressing the presentation of hierarchical data ,e.g. syntactic search, and use of additional layers, e.g. named entities.

TEXT Version 3.0 and beyond, cont

The future development of CLARIN-FCS will be guided by popularity of the features and the cost of participation. We will try to keep CLARIN-FCS open to as many users and language repository providers as possible.

This is where your center can **contribute to the uptake of FCS 2.0**, adapting endpoints to your search engines, bringing more users to your resources by taking part in the federated content search.

TEXT

((End of part 1))