

LexFCS – Extending the Federated Content Search for Lexical Resources

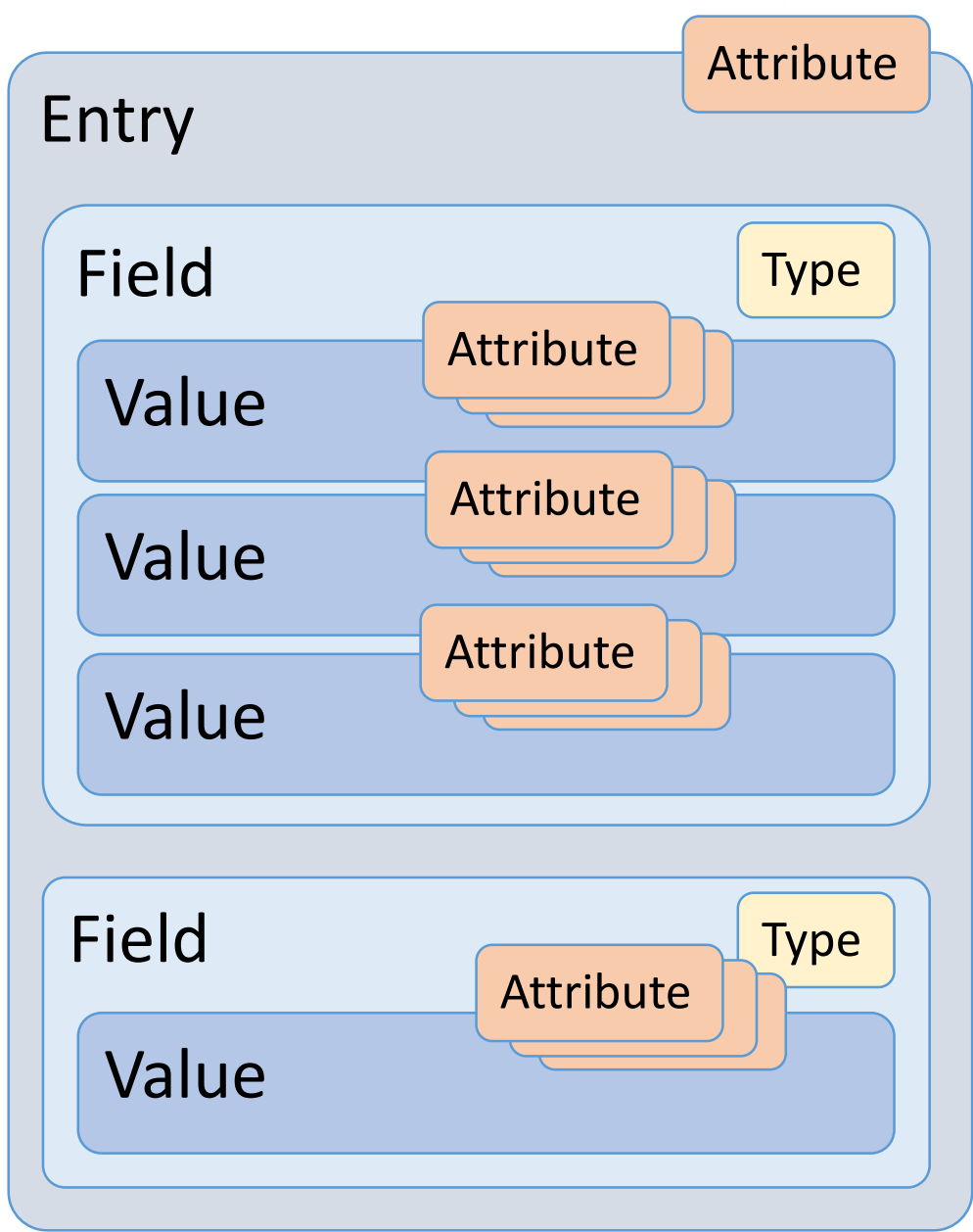
Motivation

- **CLARIN Federated Content Search (FCS)** focussed on (annotated) full-texts, corpora or similar flat text structures
→ unable to include more *structured information*, especially lexical resources like *dictionaries*, *word nets* and *graphs*
- Difficulties for querying and displaying lexical resources in flat structures without loss of information

→ **LexFCS** extension required

Data Model

- Key-value based Lemma entries
 - **Entry:** a *single* result, with optional language info
 - **Field:** set of values grouped by **type**, e.g.,
entryId, *lemma*, *phonetic*, *translation*, *transcription*, *definition*, *etymology*, *case*, *number*, *gender*, *pos*, *segmentation*, *sentiment*, *antonym*, *hyponym*, *hypernym*, *meronym*, *holonym*, *synonym*, *subordinate*, *superordinate*, *related*, *ref*, *senseRef*, *citation*
 - **Value:** actual “content”/value with attributes for additional context, e.g.,
xml:id, *xml:lang*, *langUri*, *preferred*, *ref*, *idrefs*, *vocabRef*, *vocabValueRef*, *type*, *source*, *sourceRef*, *date*
- Natural serialization in **Lex Data View** (XML)



Contributions

- (backwards) compatible extension of FCS for lexical resources,
 - “lexical resources” capability to indicate endpoint support,
 - **LexCQL** – query language based on *Contextual Query Language* (OASIS/Library of Congress) with constraints for searchable fields, operators and relation modifiers,
 - **Lex Data View** – key-value based result serialization, (optional) **LexHits Data View** as extension of *BASIC Hits Data View* with inline annotations.
- First prototypes in Text+ including client and endpoint implementations, custom visualization in FCS Aggregator

Query Language

- Based on *Contextual Query Language* with constraints
 - Boolean operators: *AND*, *OR*, *NOT*
 - Relations: *=*, *==*, *is*
with Modifiers: *unmasked*, *lang ignore/respectCase*, *ignore/acceptAccents*, *honorWhitespace*, *regexp*, *partial/fullMatch*
 - Indexes (searchable fields): based on data model, with additional entry-based language field “*lang*”
- Examples
 - **lemma** = "car" · car · "car" · "car wash"
 - **pos** = "NOUN" **AND** synonym = "house"
 - **lang** = "deu" **AND** translation =/lang=eng "member of parliament"
 - **pos is** <https://universaldependencies.org/u/pos/NOUN>

Features Overview and Demo

- Grammar with optional speech playback
- Hierarchical structure of e.g. definitions
- Highlighting of relations between values using *xml:id* and *idrefs* attributes, e.g., for definitions, examples and senses
- Citations, examples with date and source / reference
- Internationalization (user interface, with *vocab(Value)Ref* translation of POS

Next Steps

- Collecting more feedback about
 - Lex Data Model,
 - Requirements and use-cases of users and data providers,
 - User interface and visualization
- Finalizing LexFCS specification proposal with CLARIN FCS Task Force

Specification draft and examples
<https://gitlab.gwdg.de/textplus/ag-fcs-lex-fcs-dataview>



TPPSSI demonstrator
<https://tpssi-demo.saw-leipzig.de/lex/>



Nationale
Forschungsdaten
Infrastruktur



The NFDI consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG) – project number 460033370



Sächsische Akademie
der Wissenschaften
zu Leipzig

Erik Körner, Uwe Kretschmer
koerner@saw-leipzig.de