

CLARIAH Shared Development Roadmap: Data Model

Introduction

The Shared Development roadmap provides:

- ▶ A *multi-layered* overview of **CLARIAH services**
- ▶ From a generic scholarly perspective to a specific technical perspective

In order to:

- ▶ pave to road for CLARIAH-PLUS and possible successor: decide what to postpone and what not to do or postpone (**planning**).
- ▶ have a complete and transparent **overview**
- ▶ promote **interoperability**, cross-WP collaboration and **harmonize** solutions

Definitions

A CLARIAH Service:

- ▶ is initially formulated from the perspective of scholarly needs/desires: **user stories**
 - ▶ each story is as generic as possible
 - ▶ each story is as minimal as possible
 - ▶ multiple additional stories may further describe (aspects of) a service
- ▶ is not a technical concept but an abstract high-level grouping from a scholarly perspective
- ▶ enables a particular *scholarly* workflow and has one or more **implementations**:
 - ▶ an implementation consists of one or more software and data **components**, described by name and function.
 - ▶ an implementation can have multiple **instances**
 - ▶ an implementation implements the main user story and *optionally* some of the additional stories
 - ▶ can either exist already or be proposed

Example

- ▶ **CLARIAH Service:**

- ▶ **User story:** *“As a scholar, I want to search in a corpus in order to find occurrences of certain words”*
- ▶ **Implementation:** A corpus search platform consisting of software components X, Y, Z and data components A, B, C. The components form a certain workflow.
- ▶ **Instance:** A deployed form of the the implementations, hosted at a particular institute and made available over the web.

Template (1/3)

2.2.3 Corpus Search: Text & Annotation Search

(Maarten, WP3)

User story:

As a scholar, I want to perform complex searches in text collections/corpora and in the annotations on these collections **in order to** find patterns of specific (often linguistic) constructs for my research purpose.

(2) **As a scholar**, I want to view aggregated results over my results sets, such as distributions, grouped results and statistics **in order to** be able to analyse my data and identify common trends

(3) **As a scholar**, I want to provide my own text collections **in order to** have a platform that enables me to search in them.

(4) **As a scholar**, I want to search in syntactically annotated corpora (treebanks) **in order to** find linguistic patterns for my research purpose. *[this is a more specific instance of the main user story]*

(5) **As a scholar**, I want to automatically enrich my corpus with specific linguistic annotations **in order to** find linguistic patterns for my research purpose.

(6) **As a scholar**, I want uniform and rich access to a large and diverse set of corpora (possibly within a certain domain) **in order to** have a big enough data set to do searches

Template (2/3)

Implementations & Software Components

Implementation 1: INT (implements all three stories, might implement 4 in the future. Does not really implement 6)

| Component | Function(s) | Instance @Provider |
|-----------|-------------|-----------------------|
|-----------|-------------|-----------------------|

| | | |
|-----------------------------------|--|---|
| Blacklab (using Apache Lucene) | <ul style="list-style-type: none"> Storage engine for text and annotations Query & search engine Indexer to process text corpora with annotations (in specific formats) | AutoSearch@INT OpenSoNaR@INT |
| Blacklab Server | <ul style="list-style-type: none"> Web API | |
| Corpus-front-end | <ul style="list-style-type: none"> A search front-end to formulate and execute queries A results front-end to show matches in the corpus, complete with annotations An upload front-end for users to add their own data | |
| Technology Readiness Level (TRL) | Stakeholder Readiness Level (SRL) | Compatibility Level |
| 8? | | |

Template (2/3)

| | | | |
|---------------------------------|--|----|---------------------------------|
| TICCL-tools | Low-level post-OCR normalisation tools that make up the TICCL workflow. | 6 | UvT |
| Blacklab | Backend for search over large text collections, including annotations | 9 | INT |
| Corpus frontend | Generic search frontend for blacklab | 8? | INT |
| AutoSearch | Specific deployment of Corpus frontend for CLARIAH. | 8? | INT |
| GrETEL | Search in syntactically annotated corpora (treebanks) | 8? | UU |
| PaQu | Search in syntactically annotated corpora (treebanks), | 8? | RUG |
| ELAT | Collaborative web-based linguistic annotation tool (document-based, using FoLiA) | 8 | KNAW-Huc & CLST RUN (hoster) |

Figure 3: Stand-off components

Data Model

CLARIAH Services: Data model

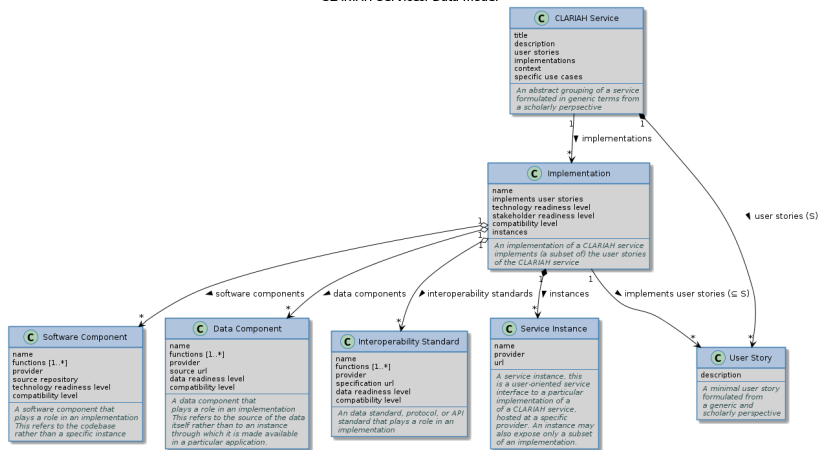


Figure 4: Data model overview