CLARIAH Heden & Toekomst: Shared Development Roadmap

Huidige Situatie

Doel: We bouwen een **shared common infrastructure** (for the humanities)

Maar. . .

- In hoeverre kunnen we daar nu van spreken?
- **Realiteit:** meerdere onafhankelijke infrastructuren
- Wat betekent het om deel uit te maken van de CLARIAH infrastructuur / CLaaS ?
  - Welke eisen stelt CLARIAH?

Huidige situatie (2)

- Gebrek aan interoperabiliteit, ook *binnen* WP3
- Gebrek aan duidelijk vastgelegde afspraken over interoperabiliteit (specificaties/requirements).
- Te weinig samenwerking tussen WP's
- Moeilijk overzicht te krijgen, veel legacy binnen WP3 en vanuit CLARIN
- Dubbel/conflicterend werk (ook binnen WP3)

Gewenste situatie

- **Harmonisatie:** Meer interoperabiliteit tussen tools en services
  - Meer samenwerking, minder dubbel werk
- Meer **transparantie**; zicht op wat er gaande is en de status -> **CLARIAH PLUS work plan**
  - WP3 Task Descriptions -> **CLARIAH PLUS task descriptions**
  - Welke use cases we bedienen -> **CLARIAH PLUS use cases**
- **Richtlijnen**: Duidelijke vastgelegde (en gecommuniceerde!) afspraken

Gewenste situatie: Duidelijke richtlijnen

- Technische richtlijnen:
  - Deelname aan de infrastructuur: **CLARIAH Infrastructure Requirements**
  - Software ontwikkeling, kwaliteit & duurzaamheid: **CLARIAH Software Requirements**
  - Verdere interoperabiliteit? CLARIAH Interoperability Requirements?
- Richtlijnen tbv gebruikerservaring
  - Gemeenschappelijke portal voor gebruikers: **Ineo**
  - branding
- Documentatie richtlijnen: **CLARIAH Documentation Guidelines**

Gewenste situatie: Tot standkoming. . .

- Interest Groups:
  - Mensen bij elkaar brengen rondom gemeenschappelijke thema's ipv werkpakketten
  - Text, AV, Annotation, LoD, DevOps, UI/UX, Workflows, ad-hoc . . . .
  - Loopt erg stroef (weinig animo/tijd)
- Technical Advisory Committee

Shared Development Roadmap: Introductie

Behoefte aan een duidelijke visie en toekomstplan over heel CLARIAH: **CLARIAH Shared Development Roadmap**

Overkoepelende doelstellingen:

- Concreet maken wat CLARIAH PLUS gaat opleveren in 2022-2023
- Richting bepalen voor eventueel vervolgproject
- Focus op generieke infrastructuur en cross-WP samenwerking
- Input voor board om budget op aan te passen

Shared Development Roadmap: Introductie (2)

Wat is de Shared Development Roadmap (SDR)?

- Een meerlaags overzicht van **CLARIAH services**
- Een **CLARIAH service**:
  - is in eerste instantie geformuleerd vanuit het perspectief van de behoefte van een onderzoeker: **user stories**
    - zo generiek mogelijk
    - zo minimaal mogelijk
  - maakt een 'workflow' van een onderzoeker mogelijk *(analyse, datatransformatie, presentatie)*
  - is hier geen technisch begrip maar een high-level abstractie
  - kan meerdere *implementaties* hebben
    - Een implementatie bestaat meestal uit meerdere software- en data-componenten (stand-off beschreven)
    - Een implementatie kan meerdere instanties hebben

Shared Development Roadmap: Doelstellingen

- **Harmonisatie** van verschillende oplossingen binnen CLARIAH; welke zijn volwassen en hebben potentie? Welke kunnen afvallen?
- **Planning**; Wat mist nog? Waar willen we heen in de toekomst?
- **Transparantie & Inzicht**; het complete beeld van generiek tot specifiek
- **Interoperabiliteit en hergebruik** van software/data bevorderen
- **Evaluatie** van CLARIAH services en componenten:
  - Technology Readiness Level (TRL)
  - Compatibility Level (CL)
  - Stakeholder Readiness level (SRL)
  - Data Readiness Level (DRL)

Shared Development Roadmap: TRL

TRL	Description	Stage
0	**Idea -** Unproven, untested and largely unformulated concept	**Planning (pre-alpha)**
1	**Initial Research -** Basic (scholarly) needs observed and reported	
2	**Concept Formulated** - Initial technology/application has been concept formulated	
3	**Proof of Concept** - Initial Proof-of-concept of key functionality . Concept presented for initial feedback from scholarly users. Not yet validated and not suitable for end-users yet.	**PoC (alpha)**
4	**Validated PoC** - Validated Proof-of-concept of key functionality. Technology validated in its own experimental setting (e.g. the lab). Not mature enough for end-users yet.	
5	**Early Prototype** - Technology validated in target setting (e.g. with potential end-users)	**Experimental (beta)**
6	**Late Prototype** - Technology demonstrated in target setting, end-users adopt it for testing purposes.	
7	**Release Candidate -** Technology ready enough and in initial use by end-users in intended scholarly environments. Further validation in progress.	
8	**Complete** - Technology complete and qualified, released for all end-users in scholarly environments.	**Production (stable)**
9	**Proven -** Technology complete and proven in practice by real users.	

Shared Development Roadmap: CL

CL	Description
A	**Excellent -** Technology adheres to as-good-as all posited infrastructure and software requirements.
B	**Good -** Technology adheres well to the requirements, there only some minor lapses
C	**Adequate** - Technology adheres to a sufficient amount of requirements, but some major ones are lacking.
D	**Lacking -** There are too many major requirements that are not met
E	**Bad -** Many requirements are not met.
F	**Unacceptable** - Technology violates or is completely dismissive of most requirements. It can not possibly be accepted without drastic changes.

Figure 1: Compatibility Level

Shared Development Roadmap: SRL

We use the **Stakeholder Readiness Level** (SRL), a measure that defines the user readiness of a new service to be used by scholars. This measure can be used for example to prioritize development using criteria such as:

- **Value**: the added value of the service for scholars (1-10)
- **Support/Commitment**: the enthusiasm in the community to adopt the service (1-10)
- **Cost**: costs for development but also cost involved for using (1-10)
- **Adaptability**: the level of adaptability in existing work processes (1-10)
- **Risks**: an assessment of the risks and their manageability that are involved in using the service (1-10)

CLARIAH Service: voorbeeld

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

CLARIAH Service: voorbeeld

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)	• 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	• Web API	
Corpus-frontend	• 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?		

CLARIAH Service: voorbeeld

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
FLAT	Collaborative web-based linguistic annotation tool (document-based, using FoLiA)	8		KNAW-Huc & CLST RUN (hoster)

Figure 4: SDR example

CLARIAH Services vanuit WP3 (1)

- OCR/HTR: *PICCL/TICCL*
- Corpus Search: Text & Annotations: *Blacklab*, *corpus-frontend*, *AutoSearch*, *OpenSoNaR*, *GreTeL*, *PaQu*
- Manuele linguïstische annotatie: *FLAT*
- Automatische linguïstische verrijking (NLP): *Frog*, *Alpino*, *UD-Pipe Frysk*, *DeepFrog*
  - voor Nederlands
  - verrijking historisch Nederlands

CLARIAH Services vanuit WP3 (2)

- Lexicon service:
  - Dialectwoordenboeken: *WLD*, *WBD*, *WGD*, *WALD*
- Dataconversie: *Piereling*, *OpenConvert*
- Spraakherkenning:
  - Voor Nederlandse dialecten: *ASTA*
  - Vanuit Stichting Openspraak: *Nederlandse ASR (oral history)*
- Audio-acquisitie in surveys: *SPAQ*
- FAIR Vocabularies

Oproep

Oproep aan alle WP3 deelnemers:

- Denk/werk mee aan de software/infrastructure requirements
- Denk/werk mee aan de shared development roadmap (deadline: 28 okt)

Links

- CLARIAH Shared Development Roadmap
- CLARIAH Infrastructure & Software Requirements
- CLARIAH Werkplan