CLARIAH-PLUS: Roadmap and Call to Action

Antal van den Bosch, Astrid Kulsdom / version 1.0 / July 16, 2020

Purpose of this document

This document summarizes the shared infrastructure that defines CLARIAH in its essence: a single *Common Lab Infrastructure* for the Arts and Humanities. Where the CLARIAH-CORE programme produced all the necessary ingredients for a common lab, they remained unconnected. The continuation of the same work packages (WPs) 1-5 and the addition of a new WP6 in the CLARIAH-PLUS programme has so far only strengthened the momentum of unconnected infrastructure building. The board has now acknowledged this as a critical point of failure for the CLARIAH-PLUS programme, for attaining its promises and goals, for its chances of formulating a successful follow-up programme and even for its chances of getting favorable midterm evaluations. Concerted action is required to mitigate this significant risk. This document therefore also proposes steps that, using the substantial resources granted to us, would allow us to change course.

The Corona pandemic of 2020 has thwarted the planned live brainstorm about these changes in the CLARIAH-PLUS programme. In subsequent online board meetings the plan has emerged to instead consult all WPs and derive an inventory of all shared infrastructures into a new Roadmap for CLARIAH-PLUS. In May and June 2020, consultancy talks were held by AvdB and AK with representatives of all WPs. This document aggregates all developed and planned technologies across WPs into shared points of interest that should be upgraded to a Roadmap of actual shared, joint Interest Groups (IGs) rather than disconnected developments.

Earlier, in brainstorm meetings in October 2018 and in the wake of the introduction and adoption of the CLaaS infrastructure, these shared points of interest were already identified as the CLARIAH-PLUS Interest Groups (IGs):

- DevOps
- Preservation
- Security & Monitoring
- Audiovisual processing Infra
- Text Infra
- Image Infra
- Geo Infra
- Annotation Infra
- Linked Open Data Infra
- Workflow
- Curation
- UI/UX

We propose to keep this inventory as the basis of our summary, with the exception of Image and Geo infrastructure, which have not emerged from the consultancy round and which can be seen as adjunct IGs developed mainly by the KNAW Humanities Cluster, and due to be connected to the CLARIAH infrastructure by WP2.

The key goal of each IG is to **research, negotiate, propose, and implement technical choices** that become a standard requirement in CLARIAH.

It may be good to recall that the CLaaS infrastructure (Figure 1) organized IGs 1-3 (DevOps, Preservation, and Security) in CLaaS' **Provisioning Services** layer, while the remainder of the IGs are placed in the **Domain Services** layer.

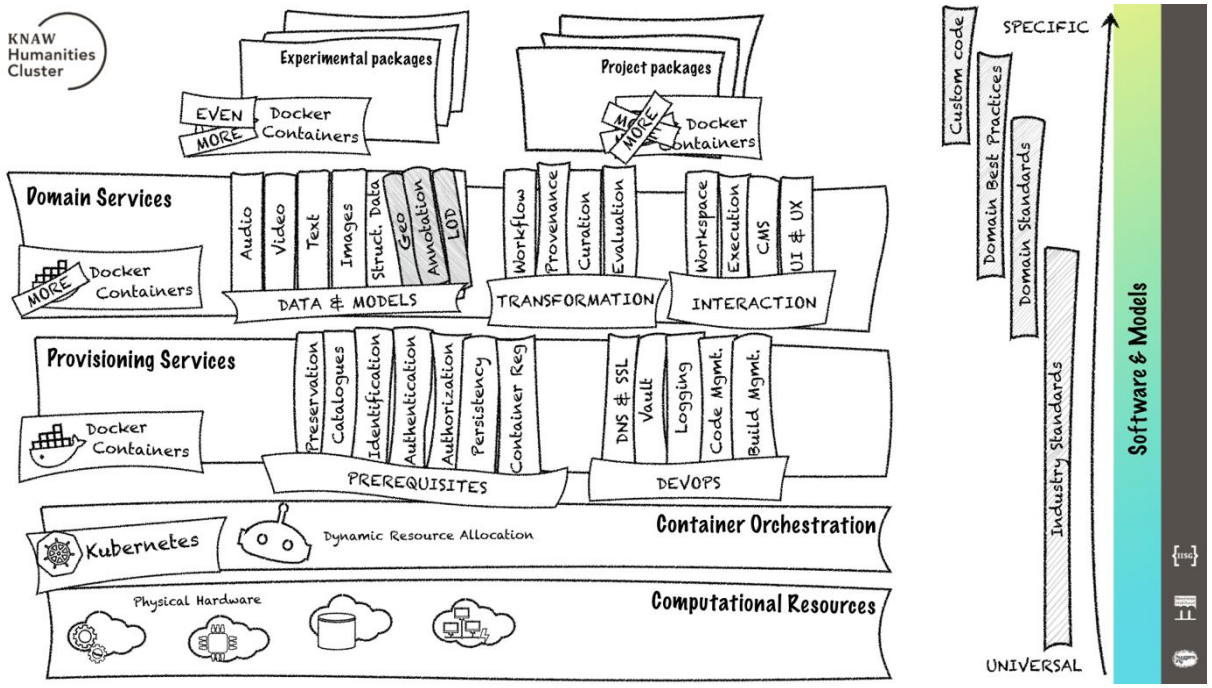

Figure 1. *The CLaaS architecture (Gertjan Filarski, C+19-137).*

This document is structured as follows. First, we aggregate the information gathered during the consultancy talks using the 10 IG aggregate labels. We then mention several additional recommendations that were suggested and brought up during the consultancy meetings on ways and methods to foster more unified collaborative work in CLARIAH-PLUS. We conclude with a call to action aimed at the board and the WP leaders to jointly decide on and act on.

Shared infrastructure: Mapping the route

All WPs have a vested interest in at least half of the IGs. WP5 is involved in 10 IGs; WP3 in 8; WP6 in 7; and WP4 in 6. Unsurprisingly, WP2 is involved in essentially all IGs, but is the logical lead for IGs 1 and 3.

For each IG, one WP is named lead and one or two members are named IG Coordinator; in the final section of this document we expand on what we mean with these labels in terms of governance and responsibility.

Disclaimer: Names mentioned in this overview are the result of a first inventory, the composition of the IGs is therefore still subject to change. This is a living document: please send us your additions and corrections and they will be updated in a next version.

1. DevOps

Lead: WP2
IG Coordinator: Lucien van der Wouw (WP2/HuC)
Members: Hennie Brugman (WP6/HuC), Martijn van de Donk (WP5/IBG), Maarten van Gompel (WP3/HuC)

Inside as well as outside WP2 several initiatives have been taken to create toolchains for packaging, configuring, virtualization, and wrapping. There is room for further integration.

WP2: Kubernetes
WP3: LaMachine, CLAM
WP5: DANE
WP6: Toolchains of INT and VU, 'tools to data' scenario KB

2. Preservation

Lead: WP6
IG Coordinator: Hennie Brugman (WP6/HuC)
Members: Enno Meijers (KB), Roeland Ordelman (WP5/IBG), Herbert van der Sompel (DANS) Jerry de Vries (WP3/DANS), Joris van Waesberge (WP2/HuC), Richard Zijdeman (WP4/IISG)

Domain-specific KOS/data models are addressed in all phases of the data life cycle, most prominently in Curation and Preservation.

WP3: CMDI as use case for preservation, in collaboration with WP2
WP2: End-user facing repositories versus long-term preservation repositories, in

collaboration with WP3, WP4, WP5, and WP6.

3. Security & Monitoring

Lead: WP2
IG Coordinator: Jan Pieter Kunst (WP2/HuC)
Members: Jaap Blom (WP5/IBG), Daan Broeder (WP3/HuC), Hennie Brugman (WP6/HuC)

Although IGs 1, 2 and 3 appear logically connected to WP2, WP5 has collaborated with WP2 on the migration of OpenConext to Satosa to create an authentication method that adheres to CLARIAH-wide requirements. WP3 mostly works with the standard CLARIN federated authentication; integration of these efforts appears warranted. A connection to the SURF Data Exchange service has been suggested in WP6.

WP2: Authentication with OAUTH
WP3: CLARIN Federated authentication; a solution for authentication across web services in a workflow is urgently needed.
WP5: Same authentication as WP2
WP6: SURF data exchange solution

4. Audiovisual processing infrastructure

Lead: WP5
IG Coordinator: Roeland Ordelman (WP5/IBG)
Members: Henk van den Heuvel (WP3/RU)

Both video and audio processing have been important resources in WP5. In particular, an existing speech recognizer for present-day Dutch has been applied to all speech collections underlying WP5's Media Suite, opening the possibility of full-text search of spoken audio fragments. In WP3, new tasks are focused on the training of speech recognizers for regiolects (e.g. Brabantish, Lower Saxon, Limburgish, Zeelandish). The partners involved (IBG, CLST and Meertens) are well-connected and also see a natural division of labor and expertise: Meertens offers curated data and represents the typical end user, CLST trains recognizers, IBG runs recognizers as services.

WP3: Speech recognition, speech alignment, speech acquisition over the web
WP5: Speech Recognition as a service, Computer vision as a service (both on bulk processing level and individual collection use).

5. Text

Lead: WP3
IG Coordinator: Jesse de Does (WP3/INT) / Maarten van Gompel (WP3/HuC)
Members: Hennie Brugman (WP2/WP6/HuC), Albert Meroño-Peñuela (WP4/VU), Roeland Ordelman (WP5/IBG)

There is a CLARIAH-wide need for robust text processing technologies that can handle historical as well as current Dutch texts. Partners like VU, INT and RU have contributed different components in WP3 and WP6. Further development should also be aimed at further integrating these efforts.

Nederlab has been one of the marriage gifts to CLARIAH. Being funded as an NWO Groot project, it has produced a portal to 18 billion words of historical Dutch texts. Nederlab offers a blueprint of a Workspace, like WP5's MediaSuite. Aside from the portal, Nederlab is based on a pipeline of WP3 technologies for the automatic analysis of historical Dutch texts.

WP2: Text repository for sharing large collections of files (such as scans of documents or xmls with data from those scans). It currently supports various serializations and versions, and it offers flexible (batch and incremental upload of documents and download via API.

WP3:

- PoS-tagging for early Dutch (INT, RU for Nederlab) and Frisian (FA)
- FoLiA (Format for Linguistic Annotations); convertors from/to FoLiA
- Frog, Dutch NLP pipeline; DeepFrog (RU, Meertens)
- Alpino, and many more

WP4: Language models for hypothesis generation. We have been training transformers (BERT, GPT-2) with large collections of papers in social economic history. The idea is to combine the learned language models with symbolic knowledge from the WP4 knowledge graphs of historical statistics, in order to suggest new hypotheses.

WP5: interested in the application of text processing tools (Tf-idf, neologism extraction, Word2vec, N-grams, Named-entity recognition, Topic modelling and word trees, sentiment analysis) on (1)subtitles or speech recognition transcripts, metadata descriptions and OCR-data.

WP6:

- Named entity parser for 17th-century Dutch (VU)
- Search interface combining 17th-century newspaper metadata with KB newspaper metadata
- Data storage environment for KB publishers' data (unencrypted ePubs)
- Apply text processing tools in a secured environment provided by KB ('tools to the data')
- Text-Fabric (DANS)

6. Annotation

Lead: WP6
IG Coordinator: Marijn Koolen (WP2/HuC)
Members: Jaap Blom (WP5/IBG), Maarten van Gompel (WP3/HuC), Roeland Ordelman (WP5/IBG)

The enrichment of various types of data with all kinds of annotations is a shared interest of several WPs, although requirements may differ greatly depending on the material to be annotated. The further development of a network of distributed annotation servers in WP2 calls for close collaboration with WP3, WP5 and WP6, who have defined several annotation-related tasks.

WP2: Scholarly Web Annotation Server
WP3:

- Manual annotation of (federated) search results, annotation of elements in a structure that is part of search results by a query.
- FLAT: FoLiA Linguistic Annotation Tool

WP5:

- Media Suite multimedia annotation tool
- CLaaS integration with ELAN, optionally CLaaS services integrated in ELAN.

WP6: INCEpTION, a semantic annotation platform

7. Linked Open Data

Lead: WP4
IG Coordinator: Albert Meroño-Peñuela (WP4/VU) / Richard Zijdeman (WP4/IISG)
Members: Willem Melder (WP5/IBG), Jauco Noordzij (WP2/HuC), Menzo Windhouwer (WP2/HuC)

The large amount of structured datasets in CLARIAH calls for a shared approach to creating, publishing, and facilitating access to them. WP4 and WP2 employ the Linked Open Data paradigm to represent data and metadata, while WP3 prefers CMDI and offers conversion to RDF. WP2 and WP4 have developed parallel LOD solutions; a next step would be to synthesize these efforts. The LOD Interest Group also collaborates closely with IG 6, as LOD forms the foundation for web annotations.

WP2: Anansi
WP3:

- CMDI2RDF
- FoLiA Set Definitions
- Codemeta in LaMachine

WP4: Druid
WP5: DIVE (deprecated). Media Suite aims for a hybrid approach where LOD can

provide an auxiliary mechanism for connecting information sources.

8. Workflow

Lead: WP3
IG Coordinator: Daan Broeder (WP3/HuC) / Maarten van Gompel (WP3/HuC)
Members: Jaap Blom (WP5/IBG), Hennie Brugman (WP6/HuC), Martijn van de Donk (WP5/IBG),Jauco Noordzij (WP2/HuC), Roeland Ordelman (WP5/IBG), Joe Raad (WP4/VU)

WP2 works to facilitate modular distributed workflows that bring together components developed within and outside of CLARIAH to facilitate the different steps of research processes. Workflows used in WP3, WP4,WP5 and WP6 contain potential components at all levels. This Interest Group has a strong connection to IG 1 (DevOps), which focuses on toolchains that support workflows.

WP2: Provides authentication tooling, aims to support entire workflows with a modular approach

WP3:

- LaMachine
- PICCL
- CLAPOP
- VU Reading Machine (VU-RM) Pipeline
- Search in linguistic data: PaQu, GrETEL 4, OpenSoNaR, MIMORE, TDS, AutoSearch, Nederlab and several others originating from CLARIN-NL

WP4:

- COW for creating knowledge graphs, publishing (in DRUID or other SPARQL endpoints), accessing via
  Data Stories (online papers with visualisations) and GRLC (to store and share (SPARQL) queries and create API's on top of them)
- Datalegend

WP5 :

- DANE (AV processing pipeline, see also AV infrastructure)
- search and analysis options in the Media Suite, multiple visualization options

WP6: Several use cases apply workflows, also workflows where several partners are involved and data has to be converted and passed on.

9. Curation

Lead: WP1
IG Coordinator: Sebastiaan Derks (WP1/HuC)
Members: Willem Melder (WP5/IBG), Ruben Schalk (WP4/UU), Mari Wigham (WP5/IBG)

Several WPs develop curation tools that enable the conversion between data and metadata standards. A Data Officer was appointed specifically by CLARIAH to oversee this joint effort, also with an eye on sustainability (and with a link to the Preservation IG).

CLARIAH would benefit from converters between data formats. Converters are developed across WPs, planned and unplanned, and many other converters are developed elsewhere; a converter toolbox would be an asset for CLARIAH.

WP2: Many collections have been converted to linked data format (triples) and stored in Anansi, in particular for Persons, Concepts (DIAMANT) and Locations.
WP3: Data curation by RU Nijmegen. CMDI Metadata creation and curation at HuC. Metadata creation and curation for tools /service /application by UU.
WP4:

- The bulk of the work in WP involved converting existing data into RDF format (triples). Special tools were developed to that end, among them COW, CATTLE, and the currently obsolete QBER.
- Recommendation for vocabularies used in the humanities based on various projects such as Golden Agents and CLARIAH user days: https://github.com/clariah/awesome-humanities-ontologies

WP5: Linked data, vocabularies.

10. UI/UX

Lead: WP6
IG Coordinator: Bas Doppen (WP6/HuC)
Members: Sebastiaan Fluitsma (Communication/WP1/HuC), Bram van den Hout (WP4/IISG), Roeland Ordelman (WP5/IBG), Paul Trilsbeek (WP3/MPI)

Sometimes characterized as CLARIAH's Achilles heel, much of CLARIAH's success in outreach depends on attractive and effective UI/UX. As yet there is no concerted action, but much work has been carried out nonetheless in different WPs. Most of the portals and search engines in CLARIAH, and also in Nederlab, have been equipped with standard (Google-like, KWIC, timeline, word cloud, network, clustering, …) visualizations using many different standard and custom libraries. Should best practices be shared? Should there be a more unified CLARIAH UI/UX?

Efforts in this IG link naturally to the communication plan developed in 2020 by Sebastiaan Fluitsma for an integrated CLARIAH portal/website.

WP3: Search result and annotation visualization: OpenSoNaR (corpus composition), MIMORE (geomaps), FLAT (various linguistic structures), PaQu, GrETEL (syntactic structures), Nederlab (various)
WP4: Druid's Table & Browser, Yasgui, Data Stories

WP5: Media Suite, including Data Stories
WP6: Custom interfaces to TextFabric

Recommendations

The following additional recommendations have been formulated during the consultation round, usually unsolicited, and triggered by the general question whether there were obvious points for improvement that could benefit the entire CLARIAH-PLUS programme.
We should have:

- A more concrete focus on **education and training (CLARIAH LEARN)**: budget for the creation of tutorials, organize a *call for teaching pilots* or request at least one deliverable aimed at training (e.g. video tutorial, worksheet) within Research Pilots;
- A **central location and overview of**
    - shared documentation, tutorials, and best practices;
    - gold-standard data;
    - CLARIAH software components and workflows to aid our own developers and advanced users

We should promote:

- Jupyter Notebooks/literate programming to further involve interested researchers;
- Efficient and professional standards for reporting.

We should organize:

- More in-depth meetings to showcase what's happening in WPs/use cases in order to stimulate knowledge exchange.

A Call to Action

The CLARIAH board and WP leaders should commit to the IGs as proposed above, allocating WP resources to these projects. In the majority of cases, resources that have already been allocated for tasks relating to these IGs can be pooled straightforwardly from the contributing WPs, as many of their tasks were already allocated to the topics at hand, but the content of the work should at least be partially refocused on integration of efforts.

IG teams should be formed by all developers across WPs who actively work on the IG's topic. One of the IG team members should be assigned the 'IG coordinator' role[1]; the primary task associated with this role is to **coordinate the work in the IG**. All IG coordinators are members of the CLARIAH Technical Committee. The WP leader who is assigned the 'lead' role in the list of IGs proposed above is in charge of **monitoring**

[1] Names of IG co-ordinators were already suggested; many of the suggested people have agreed to take on the role. Not all vacancies have been filled.

progress of the IG, and is in direct touch with the IG coordinator (e.g. through frequent meetings). The IG coordinator is in charge of **drawing up a working plan for the IG** (which, naturally, links with and is dependent on the plans of the different WPs involved), and is responsible for **formulating and proposing technical choices**. The latter are discussed in the Technical Committee (and, if they require significant reallocation of resources, in the Board); the former are discussed with the WP Lead (and again, if they require significant reconsiderations in planning, discussion is brought forward by the WP Lead in the Board, who can invite the IG coordinator in the Board meeting to clarify the issue at hand).
