**Visiedocument voor WP6 van CLARIAH+**

*V3 Karina, 8 maart 2021*

**Inhoud**

* Globale visie voor WP6
* Werkplannen per partner
  + HuC versie maart 2021
  + KB versie maart 2021
  + VU samengevat op basis van versie 2019
  + INT samengevat op basis van versie 2019
  + DANS (nog toe te voegen later dit jaar
* Ideeën voor toevoegingen aan plannen voor WP6

**Globale visie voor WP6**

CLARIAH Plus bouwt een digitale infrastructuur voor de Geesteswetenschappen. Werkpakket 6 concentreert zich hierbinnen op data en tools voor het onderzoek van teksten. Beoogde onderzoekers zijn literatuurwetenschappers, historici, religiewetenschappers, filosofen, etc.

Binnen de infrastructuur zal WP6 informatie aanbieden die relevant is voor tekstonderzoekers en voor ontwikkelaars. Welke data zijn beschikbaar, welke tools zijn beschikbaar, welke documentatie en welke instructie is beschikbaar? Bij wie moet je zijn voor welke data en welke expertise en hoe kun je ervoor zorgen dat nieuwe data, tools en expertise zichtbaar wordt binnen de CLARIAH-community? De nieuwe website die binnenkort wordt opgeleverd zal een belangrijke rol spelen in het centraal beschikbaar stellen van informatie. (**Coördinator**: HuC (Huygens): Management; Dissemination and teaching)

Tekstonderzoek omvat een scala aan werkzaamheden die worden uitgevoerd met een grote hoeveelheid methoden en technieken. Niet alle data en alle tools nodig voor die werkzaamheden zullen al daadwerkelijk in de infrastructuur kunnen worden ondergebracht. Informatie delen en het vindbaar maken van momenteel beschikbare data en tools is de eerste stap. Waar we in een volgende fase van CLARIAH op uit willen komen is dat onderzoekers voor elke stap in hun werk gebruik kunnen maken van beschikbare tools en data, en dat de resultaten uit hun werk het juiste formaat hebben om andere tools op los te laten zonder verlies van eerder toegevoegde metadata en annotaties. Zo’n super-workflow waarin ook gebruikers die niet-code-vaardig zijn hun weg kunnen vinden is nu nog te idealistisch, maar we houden voortdurend in gedachten welke stappen we moeten doen in die ideale richting. Door instructie te geven hoe specifieke tools kunnen worden gebruikt kunnen we erachter komen wat er nog nodig is om die tool in de pipeline-van-de-toekomst in te passen. We zullen laten zien welke tools voor welke aspecten nu de beste keus zijn en inventariseren waar nieuwe wensen ontstaan.

WP6 stelt zich ten doel om zoveel mogelijk tussenstappen zichtbaar, beter en werkbaarder te maken; we bespreken met elkaar welke daarvoor in aanmerking komen en bekijken of het haalbaar is. Een voorbeeld: Transkribus is nu op een niveau dat het een goede knoop in de pipeline zou zijn. Wat kunnen we daarvoor doen? En zijn er vergelijkbare knopen?

Veel moet nog worden uitgeprobeerd en kan pas een plek vinden in de toekomstige infrastructuur. De huidige infrastructuur is dus ook een virtuele infrastructuur die ons helpt na te denken over de toekomstige infrastructuur. En daarin zijn de vragen en wensen van onderzoekers van tekstuele bronnen steeds leidend.

Een voorbeeld van de mogelijke stappen in het onderzoek naar teksten:

* beschikbaarstellen van hoge kwaliteit scans van handgeschreven manuscripten of gedrukte documenten
* HTR of OCR
* Het maken van een digitale wetenschappelijke editie van een tekstuele bron (zorgvuldig gecontroleerde tekst, voorzien van annotaties in allerlei categorieën, met een inleiding over ontstaan, auteurschap, provenance, historische context etc.)
* Lemmatisering en part-of-speech tagging om tekst beter te kunnen analyseren op concepten bijv.
* NER en NEResolution om teksten in historische context (netwerken) te kunnen plaatsen, als deel van het semantisch web
* Formuleren van een inhoudelijke onderzoeksvraag
* Toepassen van tools die kunnen helpen bij het beantwoorden van de vraag (bijv. auteurschap, positionering in een literaire stroming, positionering in een bepaald politiek of maatschappelijk debat, etc.)
* Schrijven van wetenschappelijke artikelen waarin de resultaten worden geïnterpreteerd ter beantwoording van de onderzoeksvraag
* Beschikbaar stellen van data en tools t.b.v. herhaalbaarheid en controleerbaarheid
* Nieuwe tool toepassen op eerdere data of onderzoeksresultaten
* Heroverwegen eerdere interpretaties
* Schrijven van een nieuw wetenschappelijk artikel
* Beschikbaar stellen van data en tools t.b.v. herhaalbaarheid en controleerbaarheid
* Etc.

**Vragen die we moeten beantwoorden:**

Welke stappen in de super-workflow-van-de-toekomst zijn nu al beschikbaar en welke nog niet?

Welke use cases hebben we nodig om lacunes in te helpen vullen?

Wat is haalbaar in de resterende tijd van CLARIAH Plus?

Hoe gaan we de issues prioriteren?

Welke issues doen we in WP6, en voor welke hebben we andere werkpakketten en de interest groups nodig?

**Werkplannen per partner**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Partner** | **Task / Deliverable** | **Category** | **Due Date** | **Status** |
| HuC (Huygens) | Roadmap for planned use cases with scholars | Documentation | 2020Q1 |  |
| HuC (Huygens) | Dissemination and teaching | (Overseeing and organizing) |  |  |
| HuC (Huygens) | WP6 Management | (Overseeing and organizing) |  |  |
| HuC (DI) | Roadmap for development of Nederlab in CLAAS | Documentation | 2020Q1 |  |
| HuC (DI) | Nederlab in CLAAS | Software | 2022Q4 |  |

**HuC (Huygens): Management**

**HuC (Huygens): Dissemination and teaching**

**HuC (Huygens): Scholarly users collaborating with partners (use cases).**

* Use case 1 “VOC”  
  The use case was designed around detecting shifts in perspective on VOC policy over time. Eventually it has been down scaled to the development of a NER parser that facilitates automatic annotation of relevant entities in de Generale Missieven, and working towards a solution by which data and annotations can be moved through a pipeline without loss of (already added) information.
* Use case 2  
  Originally conceived as a restart of a GUI for the corpus of 17th century newspapers and attached metadata, the use case has been upgraded to a CLARIAH Show Case, including a geographical component (to locate news reporters). There is great interest in a pipeline to process the raw scans from the KB: Scans -> HTR -> Correction and Metadata Annotation GUI -> Show Case Frontend GUI. Especially a HTR solution/pipeline based on open source Calamari or Tesseract would be a greatly appreciated service on CLAAS for general use by humanities researchers.
* Use case 3  
  This use case is a pilot in the “tools-to-data” strategy to allow researchers to use distant reading (or arbitrary analysis algorithms) on IPR protected data that cannot be shared outside the National Library. The use case is in a (technical design phase). Start of work on a pilot implementation of a tools-to-data solution based on SURF’s Data-Exchange is expected for Q2 2021.
* Use case 4  
  Use case projected for a HUC text repository. On hold for the moment due to capacity shortage. > Editing Correspondences
* Use case 5  
  “Information lossless text data exchange”: a use case attacking the problem of text data being handled lossy pretty much in any research workflow (e.g. TEI-XML is delivered as input and LAF annotated text is output with most of the XML annotation stripped. The intention is to explore “common text coordinates”, “blockchain”, and “turntable” solutions. As a first expirement TextFabric has been used to replicate the two types of annotation produced in the use case 1 pipeline (TEI/NAF).

**HuC (DI): Nederlab in CLAAS:**

**Project activity 1: digital text editions and text collections**

Huygens ING is making the transition from customised digital text editions towards a more generic approach. In the 'digital text editions working group' scientific methods and their requirements for tooling are thoroughly discussed. These requirements are used as input for the development of the Docere software framework. Docere takes a generic approach to text editions but supports plug-in of customised software components for edition-specific needs.

Planning and results

The working group will deliver their *requirements* from different scholarly edition practices in 2021. *Docere* will be developed further until the end of CLARIAH Plus. We will make sure that it fits real users' needs by applying it as a front-end for a *wide range of collections and text editions*. The fact that Huygens ING is migrating its online collections to a more state-of-the-art platform is an excellent opportunity to test Docere even further.

**Project activity 2: improving and extending the Nederlab legacy**

An online research environment as provided by Nederlab is never finished: it needs maintenance, but also needs new functional extensions, more efficient workflows and more state-of-the-art hosting and deployment.

Planning and results

New or improved functionality that we have on our backlog:

* Easier and more efficient indexing workflow: this will help both collection managers and individual users to make their collections Nederlab/MTAS searchable
* Improved FoLiA documentprocessing (together with WP3)
* FoLiA document services
* Provenance services for text resources
* Jupyter notebook access to Nederlab-data
* Functional extensions driven by researchers' requests
* Upgraded hosting and deployment

**Project activity 3: tooling and infrastructure for annotation scenarios**

Annotation of texts and text collections is a recurring topic in Work Package 6 and in CLARIAH Plus as a whole. WP6 recently took a coordinating role in the CLARIAH-wide 'Annotation interest group'. This interest group focuses on manual scholarly annotation scenarios. Its' work is three-fold: prepare online text collections to make it easier to annotate them, provide annotation services to store and search annotations, experiment with small clients ('micro-frontends) that implement scholarly annotation scenarios.

Planning and results

Our aim is to make *collections* of the following media types 'annotatable': text, image, audiovisual, and we intend to provide *tooling* for others to prepare their collections as well. We will implement a *generic annotation service*, based on existing standards and preferably using existing technology. We will present a number of *'micro-front-ends*' online, as examples for others of how to use this annotation infrastructure.

This activity will be a collaboration between Huygens ING, HuC-DI and other partners in CLARIAH Plus.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Partner** | **Task / Deliverable** | **Category** | **Due Date** | **Status** |
| KB | CLAAS installed in KB cluster | Software | 2019Q4 |  |
| KB | Roadmap for further development of KB Cluster | Documentation | 2019Q4 |  |
| KB | Enlarged and integrated corpus | Data | 2022Q4 |  |

**KB: Integrate KB (and other large) holdings (NA, TLA, Dbpedia)**

**KB: Facility to let enriched metadata flow back to the collection owners**

From the first version of the Work plan, to be discussed:

The project activities will result in setting up the basic infrastructure and delivering the services described in task 24 (integrate KB and other large holdings) and tasks 47 (facility to let enriched metadata flow back to the collection owners).

**Project activity 1: Basic infrastructure**

The work done in WP6 will comply with the generic CLARIAH-PLUS architecture vision of a common lab infrastructure that integrates a national network of computing and data resources also known as CLARIAH-as-a-Service (CLAAS) which will be developed in WP2. For this reason the WP6 services of the KB will be developed as generic functionalities available for WP6 and other work packages. The KB will work towards providing access to collections, a collaborative work space with computational resources and will explore the storage and reuse of annotations from both users and automatic processes.

Planning and results

The KB will contribute towards the CLAAS infrastructure by providing our data and activities CLAAS-compatible. We will adhere to the guidelines set up in WP2 and ensure the output of the KB is interoperable with the remaining CLARIAH infrastructure.

**Project activity 2: Service for providing data access and analysis**

For WP6 and other work packages there is a need to have access to large text datasets. The KB will provide a generic way to access text datasets in a secure and efficient way and will ensure this service is interoperable with the CLAAS infrastructure.

As a lot of data have IPR or privacy restrictions that impose limitations on their use in research, the service will be able to cope with such restrictions through an authorisation solution. Finally, the service will provide researchers with processing power to analyse the data they selected in an online workspace that is accessible for all CLARIAH researchers.

Planning and results

The KB will build a proof of concept of this service in the WP6 Use Case 3: Text as Data. This proof of concept is planned to be delivered at the end of 2022 and will be a collaboration between the KB, the HuC and Surf. The solution will be generic enough to implement in other CLARIAH institutes and will be interoperable with the CLAAS network.

**Project activity 3: CLAAS service for reuse of annotations**

An aim of CLARIAH-PLUS is to significantly extend facilities to enrich data with all kinds of annotations, made by both users and computational processes. As the KB provides a large number of collections where annotations can be useful to analyse sections of texts or collaborate with other researchers with similar interests the KB will actively engage with the SIG Annotations to examine ways of creating, storing and accessing annotations on text collections within WP6, fitting the annotation infrastructure that is to be developed within WP2.

Planning and results

KB will join a collaboration with the Annotations SIG where we will explore the contribution of text-based annotations within the planned annotations infrastructure with a persistent means of referencing a section of text in its various compositions (i.e. letter, word, line, sentence, etc.). This annotation infrastructure will be taken into account within the Use Case 3, as it provides a good use case for development, but will not be in scope within the pilot project. The KB will produce a research report on supplying persistent identifiers for the content of its collections.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Partner** | **Task / Deliverable** | **Category** | **Due Date** | **Status** |
| VU | Multiple smart analysis tools still to be prioritized | Software | One each year; the selection of the tools will be done in WP6 meetings with users. |  |

**VU: Multiple smart analysis tools still to be prioritized, to be selected from:**

* Deep-reading: from text to structured data
* Sentiment and opinion mining/analysis
* Biography detection and modeling
* Semantic Role Labeling
* Visualizing timelines and stories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Partner** | **Task / Deliverable** | **Category** | **Due Date** | **Status** |
| INT | Web-based environment(s) for manual corpus annotation | (Infrastructure) Software |  |  |
| INT | Web site with demonstrator tools and extensive documentation | (Infrastructure) Software |  |  |
| INT | Community of researchers | (Dissemination) |  |  |

**INT for WP3:** Develop a common tag set and annotation guidelines

**INT for WP3:** Reliable metadata for historical corpus data

**INT for WP3:** Harmonize available training data and historical lexica

**INT for WP3:** Extend training data where the gaps are most painful

**INT for WP3:** Optimization and minor adaptation of existing taggers

**INT for WP3:** Integrate tools in a workflow for corpus processing

**INT for WP3:** Test the workflow on a considerable amount of data

**INT for WP6: Create a community and a shared infrastructure to enable researchers to easily upload, annotate and correct their data**

INT Deliverables for WP6:

D4 Web-based environment(s) for manual corpus annotation

D5 Web site with demonstrator tools and extensive documentation

Added task/deliverable distilled from main task: Community of researchers.

*[Still to be described: Role of DANS and TextFabric in the WP6 perspective = use case 5]*

**Ideeën voor toevoegingen aan plannen voor WP6**

* Adapting EviDENce to fit in the CLAAS infrastructure, to help researchers train new corpora to use Doc2vec concept seach, and/or develop training material for users. Use case brough in by **Susan Hogervorst (OU), Expertise: Meiert Grootes (eScience Center)**
* Overview of different NER approaches. Use case NER for literary texts: a reinvention of the Namescape taggers as developed in the CLARIN project Namescape (with a ground truth corpus available, for example as training corpus for deep learning). **Expertise: INT, Jesse de Does, Maarten van Gompel.** Request received from **Andreas van Cranenburgh (RUG)** for the Namescape data for training of a deep learning model with BERT.
* Coreference resolution and network analysis. Possible use case to be brought in by **Andreas van Cranenburgh (RUG) and Roel Smeets (RUN)**.