# T13: Digitization and conversion workflow (joint work with Wp6 use cases; revised version)

**IDs in Google Sheets:** T013

**Author:** INT/Jesse de Does

**What exists:**

*within CLARIAH:*

* PICCL web service workflow for digitization and conversion
* various conversion tools (Piereling, OpenConvert; Nederlab conversions; ….)
* several CLARIAH partners with expertise on OCR, post-correction and conversion

*in general:*

* General purpose OCR systems (Finereader, tesseract)
* Web services/applications offering document management and processing (Transkribus, Pero OCR)
* A fascinating but confusing array of modern deep learning-based tools for layout analysis and text recognition, image enhancement, ocr evaluation, ocr postcorrection (ticcl, …), ground truth production …
* Cf. also <https://www.digitisation.eu/tools-resources/tools-for-text-digitisation/>

**What must be adapted / extended / created anew:** Infrastructure for digitization to support humanities researchers wanting to digitize or working with noisy data (eg. OCR’ed material). By infrastructure we mean a combination of tools, services, training material and support.

Researchers should be supported to

* Choose an optimal combination of tools for their data, which implies
  + Evaluation platform enabling researchers to quickly assess the potential of tools for their data
  + Ground truth creation
* Have the possibility to choose from:
  + A web service based architecture
  + Offline availability of a wide range of tools for processing large volumes of data

Researchers should be able to acquire information:

* on how to build their own corpus
* on available tools, other than the ones developed by partners
* on options for digitisation by service providers

INT can rely on work done in projects like IMPACT, SUCCEED and Transcriptorium, and is involved in the IMPACT Centre of Competence in digitisation (founding member and leading the EB). RU: PICCL.

**Revising the planning**

In response to the rapid change of the digitisation landscape and to the requirements from CLARIAH wp6 use cases, we revised the planning.

We observe a very promising but, for users, confusing proliferation of next-generation tools, partly experimental, based on deep learning which outperform the traditional methods by a large margin on difficult documents.

Developing a monolithic workflow system based on the mainstream systems is therefore suboptimal; incorporating all relevant state of the art tools is not feasible, given the different input-output requirements and the continuous evolution of the tools.

We therefore opt for a looser approach, possibly based on/connecting to the similar workflows developed in the OCR-D project (<https://ocr-d.de/en/about>).

Deliverables and planning

Originally envisaged deliverables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Taak | ID | D or M | Category | Description |
| T012 Digitisation & Conversion Infra | T012D1 | Deliverable | Software | Enhanced version of TICLL (RU)  The planned extensions to PICCL are described in the detailed project plan. |
|  | T012D2 | Deliverable | Software | Enhanced PICCL workflows  The planned extensions to PICCL are describe in the detailed project plan. |
|  | T012D3 | Deliverable | Software | Enhanced and user-friendly PICCL interface  Create a new web-based front-end for the PICCL workflow system, enabling users to  – Bypass advanced options with the help of a very simple interface (cf. PhilosTEI) that determines more options from the input supplied by the user  – Enable simple evaluation scenarios, with uploaded ground truth or gold standard  – Provide a simple demonstrator platform that allows users to assess easily how tools might work on their material |
|  | T012D4 | Deliverable | Website | Web site with demonstrator tools and documentation |

Revision of deliverables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Taak | ID | D or M | Category | Description |
| T012 | T012D1 | Deliverable | Software | Documented workflows for OCR and conversion of Dutch documents |
|  | T012D2a | Deliverable | Data | Ground truth material for Dutch-language OCR   * 17th century newspaper set * set based on conversion/enhancement to line level of the IMPACT Dutch ground truth * As much as we can grab from other sources |
|  | T012D2b | Deliverable | Data | Trained models for OCR of Dutch documents |
|  | T012D2c | Deliverable | Data+software | Evaluation of OCR tools on ground truth data: data, evaluation workflow, document on evaluation |
|  | T012D3 | Deliverable | Website | Web site with demonstrator tools and documentation |

Original INT workload planning within WP3 (needs revision)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Taak** | **Type personeel** | **Available** | **2019** | **2020** | **2021** | **2022** | **2023** | **Summed** |
| T012 Digitisation & Conversion Infra | Linguist | 0 |  |  |  |  |  | **0** |
| T012 Digitisation & Conversion Infra | Developer | 4 |  | 1 | 1 | 2 |  | **4** |
| T012 Digitisation & Conversion Infra | Computational linguist | 2 |  | 2 |  |  |  | **2** |

Partial revision:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Taak** | **Type personeel** | **Available** | **2019** | **2020** | **2021** | **2022** | **2023** | **Summed** |
| T012 Digitisation & Conversion Infra | Linguist | 0 |  |  |  |  |  | **0** |
| T012 Digitisation & Conversion Infra | Developer | 4 |  | 1 | 4 | 2 |  | **6** |
| T012 Digitisation & Conversion Infra | Computational linguist | 0 |  | 0 |  |  |  | **0** |

*!The current proposed planning concerns the INT part of the work only; the work of the other partners also needs to be reconsidered.*

The original project plan states:

INT and KNAW HuC Meertens Institute and HuC Digital Humanities Lab, Radboud University Nijmegen (RUN, CLST), Tilburg University (TiU). HuC in tandem with RUN will concentrate on backend (TICCL, PICCL, converters, software distribution in a.o. LaMachine) and INT willconcentrate on frontend, converters, information and hosting.

**Targeted/Actual users:** Digital humanities researchers depending on corpus material in an appropriate format for their research  
**Actual use (quantify!):** We are currently unable to quantify. Envisaged uses are at least the CLARIAH use cases on historical newspapers and on corpus-based history of ideas.

**Social Impact** **(concrete examples):** Researchers will be able to build their own corpora, thus facilitating their research

**Estimate in PMs (try to justify): 6**

**Lead + PMs:**

**Participants + PMs:** INT: 6