# Topic: Digitization and conversion workflow

**IDs in Google Sheets:** T013

**Author:** INT

**What exists:**

*within CLARIAH:*

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

*in general:*

* https://www.digitisation.eu/tools-resources/tools-for-text-digitisation/

Image enhancement, ocr, htr, ocr evaluation, ocr postcorrection, ground truth production ...

**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 enabled 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.

**Targeted/Actual users:**   
**Actual use (quantify!):**

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

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

**Lead + PMs:** XX:##

**Participants + PMs:** INT, RU