Workflow for Digitization and Conversion (T012/T013)

INT, KNAW Institute Meertens, RUN

July 1, 2021

Abstract

We propose an enhanced workflow infrastructure for digitization and conversion 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 an evaluation platform enabling researchers to quickly assess the potential of tools for their data as well as the suitability of their data for their research.

Users should have the possibility to choose from:

- A web application with a simple interface
- An advanced interface enabling processing options
- A web service based architecture, enabling to run both workflows and individual tools
- Offline availability of a wide range of tools in command line mode in a workflow system for processing large volumes of data

Users should be able to acquire information:

- on how to build their own corpus
- on OCR output formats and how to post-process them
- on available tools, also including tools not developed by partners
- on options for digitisation by service providers

1 Introduction

1.1 Impact

Why important for CLARIAH (scientific impact) The workflow for digitisation will take the current PICCL pipeline, powered by Nextflow, as a starting point and LaMachine (T098) as a basis for distribution. The new workflow system, with extended options and new and enhanced user interface, will more comprehensively allow also non-technically inclined Digital Humanities researchers and social scientists to build, curate, analyse and search online their own particular corpora.

Targeted/Actual users Historical and variational linguists, historians, digital humanities scholars and social scientist researchers working with (historical) text and in need for a solution to digitize or convert their data. It is difficult to quantify the use of historical annotation tools. One needs to distinguish the smaller group of corpus builders from the larger group of corpus users. Both benefit from an improved infrastructure, directly and indirectly. Obviously,

- There are many researchers using historical text
- Easy accessibility to tools may increase the number of corpus builders

Actual use Nederlab developers at RUN and INT, group of philosophers around Arianna Betti's VICI project 'e-Ideas' (UvA), 4 CLARIAH Pilot projects (UU, IISG, RUN). Prospective user: NWO VC project 'Chronicling novelty. New knowledge in the Netherlands, 1500-1850' by J. Pollmann and E. Kuijpers.

Social Impact Researchers will be able to build their own corpora, thus facilitating their research

1.2 What is available

1.2.1 Workflow systems

- PICCL (Philosophical Integrator of Computational and Corpus Libraries)
  - Web application (generic CLAM interface)
  - Web service (RESTful, powered by CLAM)
  - Offline tool chain and distributed / parallelized workflow system (orchestrated by Nextflow)

1.2.2 OCR Systems

- Tesseract, integrated in PICCL

1.2.3 Postcorrection system

- Multilingual TICCL (Text-Induced Corpus Clean-up), integrated in PICCL

1.2.4 Linguistic annotators

Ucto has been integrated into PICCL.

- Ucto provides tokenization for a wide range of languages.

Frog has been integrated in PICCL.

- Default Frog annotations in PICCL:
  - lemmatisation
  - POS-tagging
  - Named-Entity Recognition
- Optional Frog annotations in PICCL:
  - Morphological Analysis
  - Chunking
  - Dependency Parsing

1.2.5 Converters

Many are available

- Pandoc for multi-format document conversion tool
- Oxgarage for multi-format TEI-centric conversions
- Plain text to FoLiA via ucto (tokenizer)
- OpenConvert conversions, a.o. from MS word and epub to TEI
- Nederlab converter from ABBYY XML to "editorial" TEI
- FoLiA Utilities (https://github.com/LanguageMachines/foliautils)
  - ALTO to FoLiA / TEI
  - hOCR to FoLiA
  - Page XML to FoLiA
- FoLiA Tools (https://github.com/proycon/folia)
  - ReStructuredText to FoLiA (very powerful when combined with pandoc! Allows for Markdown / MSWord / LibreOffice / LaTeX / HTML→ReStructuredText→FoLiA)
  - FoLiA to plain text or ReStructuredText
  - FoLiA to columned output (tsv or CONNL-like)
  - FoLiA to HTML (for visualisation)
  - Alpino XML to FoLiA

- NAF to FoLiA and FoLiA to NAF (a CLARIAH Core collaboration between RUN and VU, still in early stage of development)
- Various partial converters from TEI to FoLiA (Nederlab DBNL converter in PICCL, Nederlab Huygens converter)

1.2.6 Software distribution

- LaMachine (T098), a meta-distribution for NLP software that allows deployment of PICCL and all dependencies.
  - as a Virtual Machine
  - as a Docker container
  - as a local installer (in a virtualenv)
  - as a remote provisioner

1.2.7 Centre of competence

The Impact centre of competence[1] provides

- A wide range of digitization tools
- A demonstrator platform allowing users to test tools

1.3 What must be created

1. More tool choices, e.g. for OCR engine (choice between Tesseract 3 and 4, possibly a wrapper for Finereader) or OCR engine output formats (e.g. ABBYY SDK)
2. TICCL
   - Configuration settings for TICCL enabling a sensible choice of background material depending on the material to be processed (i.e. intelligent sample selection), profiting from the developments in Nederlab (corpus material) and CLARIAH / eScience project TICCLAT
   - Improved implementation of word embeddings as a feature for the TICCL correction candidate ranking tool
3. More robust approach to TEI/FoLiA conversion. Current tools work for a subclass of TEI documents that is not clearly defined. Documents outside this subclass may result in invalid FoLiA. This is unsatisfactory. Two approaches could be elaborated:
   - Strict definition of the set of feasible documents for a fine-grained conversion which can be converted in a "lossless" way. Documents in this set can be converted back to TEI, thus enabling TEI input and output for workflows.
   - A possibly lossy more generic conversion for the other documents
4. Conversion from ABBYY XML to FoLiA and TEI
5. 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
6. Visualization and interactive options
   - FoLiA output of tools will be made to seamlessly interact with FLAT (T062) for further manual linguistic annotation and visualisation.
   - TEI viewer

[1] https://www.digitisation.eu/

2 Work plan

2.1 Disclaimer

The authors of the above list and the proposed partners in this and further related and linked subprojects wish to stress that the list of extensions and tools to be integrated into PICCL is not necessarily exhaustive nor hewn in rock. The world is in a constant flux, today's alpha tool may well be tomorrow's gamma choice. PICCL should evolve accordingly and keep offering researchers the best possible options.

2.2 Involved Partners

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 will concentrate on frontend, converters, information and hosting.

External collaboration will involve the IMPACT Centre of Competence.

2.3 Deliverables

ID	Description	Who	PM
D1	Enhanced version of TICCL	Meertens Institute / Tilburg University (Martin Reynaert)	6
D2	Enhanced version of PICCL, offering more options for OCR and conversion and incorporation of evaluation tools	INT (TEI conversion), Meertens, RU (FoLiA, Nextflow scripts)	8
D3	Enhanced and user-friendly PICCL interfaces	INT	6
D5	Web site with demonstrator tools and extensive documentation	INT, Meertens, All	8

Appendix I: Overview of proposed extensions to existing PICCL work flow

This table is meant to give an overview of existing CLARIAH-Core PICCL functionalities and new functionalities to be developed in CLARIAH-Plus

The table gives the references to the separate proposals we have submitted to CLARIAH-Plus. Each of these proposals might constitute a separate project and form a stand-alone application. As we see it, there is a sizeable added value to incorporating them into PICCL: this will greatly enhance their visibility towards the community and PICCL affords them interoperability capabilities which would otherwise be either unclear or far harder to achieve to the common user.

Subservice	**CLARIAH-Core**	**CLARIAH-Plus**
CONVERSION (T012 and T013)	Conversions: – Other image formats to TIFF – Djvu file to TIFF – PDF to TIFF – ALTO xml to FoLiA xml – hOCR html to FoLiA xml – Page OCR xml to FoLiA xml – PDF embedded TXT to FoLiA xml – TXT to FoLiA xml – DOC(X) to FoLiA xml – FoLiA xml to TEI xml – FoLiA xml to TXT	– ABBYY xml to FoLiA (available) – TEI xml to FoLiA (to be developed)
OCR	Tesseract 3 or 4, depending on Linux distribution	Choice between Tesseract 3 and Tesseract 4 (based on deep learning)
OCR post-correction (T012 and T013)	TICCL	TICCL will have become TICCLAT in ADAH-call CLARIAH / eScience project TICCLAT. It will have a totally new apparatus for Dutch with more guidance to select the lexicon for a particular OCR post-correction job for a particular period in time and possibly geographical area. This apparatus needs to be integrated in PICCL.

TEXT ANALYSIS - Corpus Quality Assessment and Correction Evaluation modules (T019) - Extracting Linguistic properties (T036)	- Tools FoLiA-stats or TICCL-stats deliver ngram frequency lists and basic type/token statistics - For a list of e.g. pairs of word variant versus canonical form, the AHA application calculates the numbers of e.g. insertions, deletions, substitutions, transpositions and combinations thereof and returns a Latex-preformatted table of the statistics to the user for easy integration in e.g. the test set description part of her paper.	Integrate: – R-tools by Baayen, Stefan Evert and Th. Gries – T-SCAN – Evaluation tools by Rafael Carrasco - Impact Center ([2]), by Janneke van der Zwaan - eScience Center ([3]) and Martin Reynaert. – Reynaert's tool 'Goldie-Oldie' (deployed in CLARIN-NL project VU-DNC on 1M word tokens historical 1950s part of the newspaper corpus) for aligning OCR and Gold Standard version of the newspaper articles – The AHA-submodule is to be further developed on the basis of these text alignment, analysis and evaluation tools
LINGUISTIC ENRICHMENT - multilingual	UCTO - tokenisation	
LINGUISTIC ENRICHMENT - Dutch (currently)	FROG * Default: - lemmatisation - POS-tagging -Named-Entity Recognition * Optional: - Morphological Analysis - Chunking - Dependency Parsing	- Frog has been retrained for 13th and 14th century Dutch in project Nederlab. - Frog is to be retrained for English in another subproject proposed by RUN. Retrained Frogs are to be integrated into PICCL.

[2] https://dl.acm.org/citation.cfm?id=2595221
[3] https://zenodo.org/record/1189245#.W1GLpn58JsM

SEMANTIC WORD EMBEDDINGS (T125 and T126 and T127, linked to T039)	UCTO delivers pre-processing towards normalised text corpus (Punctuation, digits and numbers in Arabic and Roman notation, dates, etc. normalized or removed. UTF-8 enforced and / or normalized. Text lowercased if desired).	* Indexing: - Word2Vec indexer - GloVe indexer * GloVe to Word2Vec index format converter (available) * SoNaR Vector submodule tools for vector exploration / exploitation, suitable for non-interactive list work: - W2V-analogy: resolve analogies - W2V-dist: cosine distance between two words - W2V-near: N nearest semantic neighbours Output from these modules will be available within the work flow to TICCL: - for enhancing ranking of Correction Candidates - as the basis for training neural networks to solve real-word errors or confusables
SUBCORPUS SELECTION REFINEMENT Topic: Document Classification (T012, T040, T041)		Researchers may well want to draw on the rich text collections available in KB Delpher or other digital repositories. We wish to enhance their manual querying and text-selection activities and greatly reduce the total effort required. After the annotation and classification tools developed in the SERPENS CLARIAH Pilot have been made more generic, PICCL integration will follow in order to provide users with non-command line access to the classification recipe to filter articles for a specific domain or topic.
INDEXING	Autosearch	
INTERFACE	- Generic CLAM-interface - (Currently defunct) @PhilosTEI interface	- New interface to be developed by INT in accordance with PI's design desiderata

Appendix II: Provenance tracking and relation with the WP3 VRE

Plans for a WP3 Virtual Research Environment (VRE) are currently in the early stages of implementation. We seek to collaborate with this effort and prevent any unnecessary duplication. PICCL defines various workflows, powered by Nextflow, the engine which orchestrates the execution of the individual tools and optionally distributes them amongst a computing cluster if so needed. The VRE also aims to do its own orchestration, but tools in the VRE context are most often webservices, i.e. a higher level of abstraction and overhead than the PICCL/Nextflow approach.

Digitisation and conversion workflows can be made available to the larger VRE effort, albeit

that these workflows are single entities as far as the VRE is concerned. This brings some challenges for provenance tracking, which is one of the major features of the VRE. Collaboration between the two tasks is required to ensure good interoperability for this as well as other aspects.