A TEI-based publication pipeline for historical egodocuments - the DAHN project

Floriane Chiffoleau, Anne Baillot, Manon Ovide

A TEI-based publication pipeline for historical egodocuments - the DAHN project

Floriane Chiffoleau, PhD candidate at Le Mans Université and ALMAnaCH Inria
Anne Baillot, Professor at Le Mans Université (3LAM) and Researcher at ICAR UMR 5191 at ENS de Lyon
Manon Ovide, Intern at ALMAnaCH Inria

"Next-Gen TEI"
October 25 - 30, 2021 | Virtual Conference

ALMAnaCH project-team

Inria

Le Mans Université

3L.AM
Langues, Littératures, Linguistique
Le Mans Université
Université d'Angers

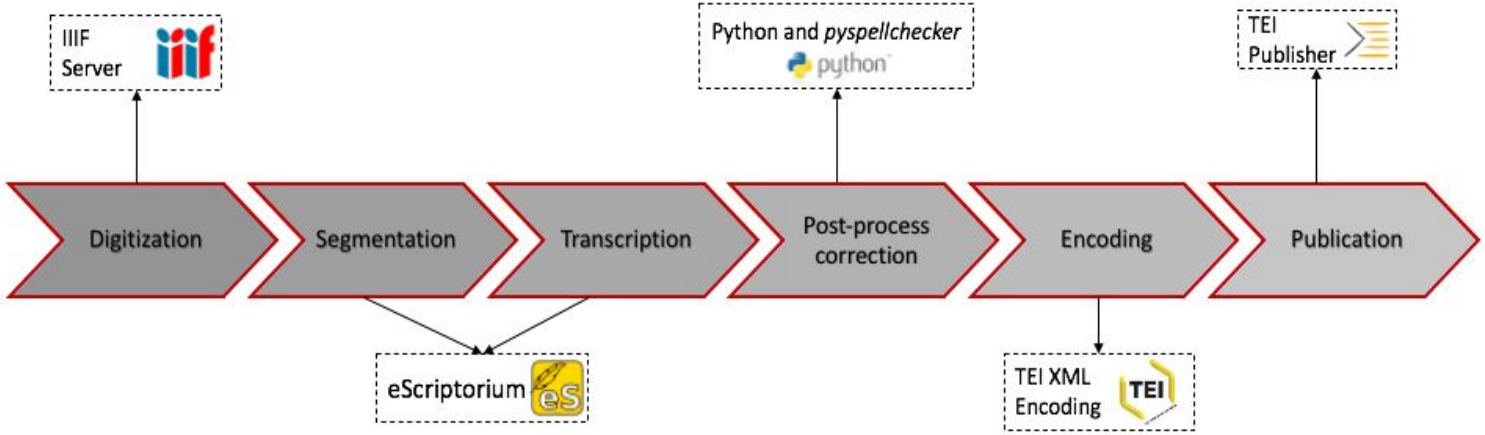
icar

Presentation

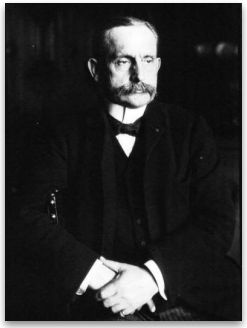
1. The **DAHN project**: a pipeline for digital scholarly edition of historical documents
2. **Experimenting** with the pipeline: working with corpora
3. Obtaining digital texts: the **eScriptorium** interface
4. Completely and accurately rendering the transcription: **encoding** with TEI XML
5. The key element of our pipeline: a TEI-based **publication application**
6. Conclusion: a **minimalist pipeline** for a maximum efficacy

The DAHN project: a pipeline for digital scholarly edition of historical documents

- Technological and scientific collaboration between Inria, Le Mans Université and EHESS
- Funded by the Ministry of Higher Education, Research and Innovation
- Work progress stored in a Github repository (DAHNProject) and documented in a Hypothèses blog (https://digitalintellectuals.hypotheses.org/category/dahn)



Experimenting with the pipeline: working with corpora



Paul d'Estournelles de Constant' corpus: development of the pipeline, segmentation/transcription, creation of models, encoding scripts, early development of the TEI Publisher application

Berlin intellectuals' corpus: updating of the TEI, new encoding scripts for modification, upload to IIIF server, counterbalance for the development of the TEI Publisher application

Publication phase tested by other corpora: Time-Us, Lectaurep (decided to develop their own instance), EHRI, Rochlitz manuscript (future corpora on our instance)

Obtaining digital texts: the eScriptorium interface

- eScriptorium: web interface for collaborative and automatic transcription projects, relying on the OCR software Kraken

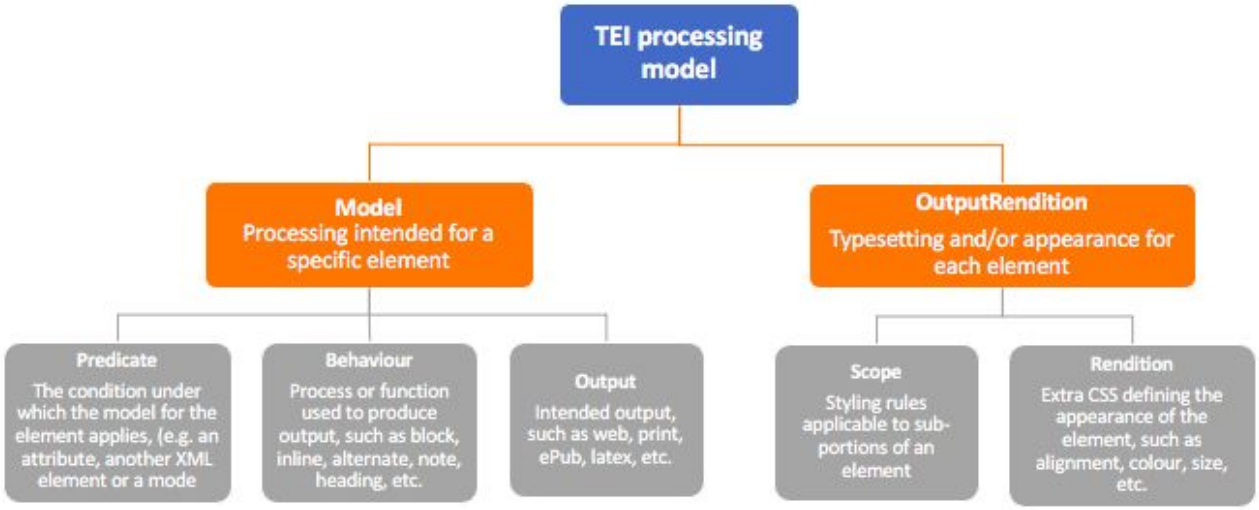
- Images import (IIIF or upload), segmentation/transcription, OCR/HTR, creation or finetune of models for segmentation/transcription, annotation of regions and lines
- Three kinds of TEI-compatible export:
  - XML Page: transformation with an XSL (https://github.com/lectaurep/page2tei)
  - Text: transformation with Python scripts
  - Future implementation: TEI export in the interface

Completely and accurately rendering the transcription: encoding with TEI XML

- TEI Guidelines
- Encoding guidelines written specifically for egodocuments
- Encoding scripts
  - Automatic encoding of the header with the generic information from the corpus
  - Encoding of the body with regular expression, using the unique architecture of egodocuments
  - Addition of few information after manual correction
  - Updating of some parts of the header and/or the body

The key element of our pipeline: a TEI-based publication application

- Developed with TEI Publisher, a tool for scholars and researchers, offering them the possibility to publish their work, without extended programming knowledge



Conclusion: a minimalist pipeline for a maximum efficacy

- The researcher has all the tools at their disposal to exploit the full potential of their corpus
- TEI ensures consistency and sustainability
- Upcoming improvement: annotation of named entities (the perfect way to exploit a corpus even more)

Contact

Floriane Chiffoleau: floriane.chiffoleau@inria.fr

Anne Baillot: anne.baillot@univ-lemans.fr

Github project: https://github.com/FloChiff/DAHNProject

Hypotheses blog: https://digitalintellectuals.hypotheses.org/category/dahn