From a collection of documents to a published edition : how to use an end-to-end publication pipeline

Floriane Chiffoleau, Hugo Scheithauer

From a collection of documents to a published edition : how to use an end-to-end publication pipeline

In 2021, during the last edition of the TEI Conference “Next Gen TEI”, I took part in a session where I presented a project I had been working on for a year and a half. This project, both relying massively on the Text Encoding Initiative and benefiting its community, focusses on the creation of a pipeline for the publication of digital scholarly editions. This pipeline, which was still a work in progress at the time of the 2021 Conference, but is now complete, aims at providing open-source, free, easy-to-use and interoperable tools; its goal is to support the editorial process from the digitization of a collection of documents to its publication in a machine-readable standard.

In the following, I will succinctly describe the six steps that compose this pipeline, and then move to the way I intend to conduct the workshop based on them.

Firstly, the collection of images that composes the corpus has to be stored and curated somewhere online, both to keep them available for researchers and for publication. For this task, we rely on IIIF, to ensure sustainability and interoperability.

The three following steps, segmentation, transcription and post-OCR correction, are performed with eScriptorium, an open-source transcription application. It offers various features: uploading images, production of ground truths, manual or automatic segmentation and transcription, using custom models, training segmentation and transcription models, to name a few. Finally, if there are any remaining errors in the transcription (in case of an automatic transcription), it is possible to either correct them manually in eScriptorium or export the files and correct them with the help of specifically designed scripts.

Once the transcription is fully done, we encode it in TEI XML. For this step, we provide various solutions, depending on the transcription file format (Page XML, XML ALTO, Text) chosen when exporting the transcription from eScriptorium. We also propose documented scripts that help automatize and speed up this process.

Encoded files are then published online with the help of TEI Publisher, an application designed for generating custom editions for corpora encoded in TEI XML. We have developed and launched a dedicated application for digital scholarly editions (DiScholEd) on this basis. It is available online together with a thorough documentation, and is conceived as an open application: new corpora can always be added to it, and we welcome new collaborations.

The goal of our workshop is to demonstrate how a corpus could be processed for publication with TEI Publisher. The workshop participants will learn to experiment with a

ready-to-use solution that provides an easy and quick publication of a corpus. They will also get tips and shortcuts to help speed up the creation of a digital edition. Moreover, by the end of the session, this workshop will provide the participants with a visualization of their respective corpus, with side by side transformed text and original image; all of which then showing what can be achieved while working with TEI in the context of an end-to-end publication pipeline.

The program for this workshop is the following: firstly, it will start with a presentation of the pipeline, its objectives and how it works. Then, the time we have will be divided into several slots corresponding to every step of the pipeline. Each slot will start with a quick presentation of what is expected of the participants and what tools they will need to use. Next, they will be allotted some time to work with their data and to process them for publication. At the end of the day, a 30mn feedback session will make it possible for each participant as well as for the workshop organizers to assess the benefits of the session and envision further possible collaborations.

Considering the number of steps in this pipeline and the time required for each of these steps, a full day is necessary for this workshop. The number of participants should be 10-15 maximum, in order for the two workshop conveners to be able to provide the necessary technical support for the hands-on parts of the workshop.

In order for the participants to be able to work correctly on the pipeline, they will need a laptop as well as the following tools: a command line interface for the execution of the scripts, an XML editor (Oxygen is the best choice) and a way to work with TEI Publisher. The latter can be launched with a local eXist-db installation, or with docker (see Documentation). An eScriptorium account will be provided to each participant. They will also have to bring their own material (textual sources preferably; images and transcription (in TXT format)) to work on (about 3 to 5 pages).

GitHub repository of the pipeline:
https://github.com/FloChiff/workshop-discholed-tei2022

Keywords: digital edition; historical manuscripts; encoding pipeline; publication workflow

Workshop leader(s)

Floriane Chiffoleau:

After a master's degree in late modern history and in "Technologies numériques appliquées à l'histoire" at the Ecole nationale des Chartes, Floriane Chiffoleau worked as a research and development engineer at Inria for a year and a half. She then started a PhD in digital humanities in October 2021 under the direction of Anne Baillot at Le Mans Université (3L.AM) and Laurent Romary at Inria (ALMAnaCH). Her research focuses on text recognition and TEI encoding.

Email: floriane.chiffoleau@inria.fr

Hugo Scheithauer:

Research and Development Engineer in the Inria ALMAnaCH team, Hugo Scheithauer holds a master's degree in art history and in "Technologies numériques appliquées à l'histoire" at the École nationale des chartes. He works on the automatic segmentation of sale catalogues for the DataCatalogue project, jointly led by Inria, the National Library of France (BnF) and the National Institute for Art History (INHA).

Email: hugo.scheithauer@inria.fr

Bibliography

Chagué, Alix, and Floriane Chiffoleau. An accessible and transparent pipeline for publishing historical ego documents. 2021. ⟨hal-03180669⟩

Chagué, Alix, and Hugo Scheithauer. 2021. page2tei, an XSL Transformation to transform PAGE XML into TEI XML (Version 1.0.0) [Computer software]

Chiffoleau, Floriane, DAHN Project, *Digital Intellectuals*, 2020-2021: https://digitalintellectuals.hypotheses.org/category/dahn

Chiffoleau, Floriane, Anne Baillot, Manon Ovide. A TEI-based publication pipeline for historical ego documents -the DAHN project. *Next Gen TEI, 2021 – TEI Conference and Members' Meeting*, Oct 2021, Virtual, United States. ⟨hal-03451421⟩

Kiessling, Benjamin et al. 2019. "eScriptorium: An Open Source Platform for Historical Document Analysis". In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2, pp. 19–19.
DOI:10.1109/ICDARW.2019.10032

Pierazzo, Elena. 2019. What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter. *International Journal for Digital Humanities*, Springer, 1, pp.1-12. ⟨10.1007/s42803-019-00019-3⟩. ⟨hal-02117714⟩