Response to Call for Papers - DH2022

Take a sip of TEI and relax: a proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition

Alix Chagué[1 2], Hugo Scheithauer[1], Lucas Terriel[1], Floriane Chiffoleau[1], Yves Tadjo-Takianpi[1], Laurent Romary[1]

1 Inria, France; 2 Université de Montréal, Canada

Abstract

We propose to present a generic workflow compatible with many different types of documents and writing systems where TEI (Text Encoding Initiative) XML plays a central role. The workflow relies on the idea that utilizing such a standard facilitates the articulation of tasks related to automatic transcription, post-processing and edition.

Proposal

Over the last decades, several breakthroughs have made the dream to automatically transcribe thousands of handwritten documents a reality (Causer et al., 2018; Sánchez et al., 2017; Seaward, 2017; Yin et al., 2013). For example, software like *Transkribus* (Kahle et al., 2017) and *eScriptorium* (Stokes et al., 2021) provide non-specialist users with simple environments to conduct transcription campaigns relying on efficient HTR[1] engines. While transposing scriptures from a piece of paper onto a text editor used to require effort and concentration, it is now possible to imagine simply pressing a button and letting your computer work while you start preparing your next cup of tea. A few minutes later, your drink is ready, and so is the transcription of the two thousand pages you needed. As automatic transcription software is about to produce huge volumes of data (Clanuwat et al., 2019; Camps, 2021. See also the Vietnamica project[2].), it seems crucial to think about how we can interact with the resulting files with maximum efficiency.

In response to previous similar initiatives (Carius, 2020), we would like to present an end-to-end workflow revolving around the use of various automatic techniques to go from a set of digital images to the actual publication of a text edition. Such techniques include, on

[1] HTR stands for Handwritten Text Recognition.

[2] Vietnamica is a research project undertaken jointly by the École Pratique des Hautes Études, the Institute of Hán-Nôm Studies, the Social Sciences Academy of Viêt Nam and the National University of Viêt Nam (Faculty of Humanities and Social Sciences). See https://vietnamica.online/

top of HTR, information extraction tools[3] and an open source and ready-to-use environment for publication. Moreover, we aim to make this framework as simple and generic as possible: it is independent from the transcription engine, and potentially compatible with any language, writing system, and any type of document (Balogh and Griffiths, 2020. See also the TEI Special Interest Group for East Asian/Japanese[4]).

Several key principles ensure the coherence of the workflow: transparency and availability of the data at each step and the use of a fully standardized format like TEI XML as the cornerstone to store all the available information. Other XML standards like ALTO[5] or PAGE (Pletschacher & Antonacopoulos, 2010) are commonly used by transcription software to export the output, but we advocate for a change of paradigm in order to give more importance to TEI earlier in the workflow (Scheithauer et al., 2020). The TEI guidelines define a set of elements to document this type of data, namely “sourceDoc” and its children[6]. Leveraging TEI from the start is essential to connect the metadata of the images[7] and documents, the text and layout information generated during the transcription, and any further editorial layer added to the raw transcription.

We imagine a configuration capable of processing a large family of TEI customizations as long as the file follows a structure (Fig. 1) in which:

- “teiHeader” stores the metadata,
- “sourceDoc” the raw transcription, and
- “body” the interpreted logical structure along with the editorial layers[8].

We thus aggregate two phases in the digitization lifecycle which are often disconnected.

Editorial operations can include preprocessing tasks such as post-HTR corrections (spell-checking) and text normalization, as well as information extraction (text mining). When the volume of data increases, extracting and linking named entities with indexes quickly risks becoming a laborious task. Instead, natural language processing tools can automate the process (Ehrmann et al., 2020; Frontini et al., 2015) all the while relying on the analysis of the sentences and words within their context. We developed *Semantic@*, a proof of concept utilizing deep learning models, to extract named entities which are then cycled back into the TEI tree (Fig. 2). The extraction of named entities (i.e. names of people, places, or dates,

[3] Rosa Stern defined information extraction as a task consisting of extracting and structuring, in semantic classes, the specific information elements contained in non-structured data for automatic processing, such as coreference resolution, relationship extraction, and named entity recognition (Stern, 2013, p. 59).

[4] See https://tei-c.org/Activities/SIG/EastAsian/ and https://wiki.tei-c.org/index.php/SIG:East_Asian

[5] See the Analyzed Layout and Text Object (ALTO) 4.2 schema specifications at https://www.loc.gov/standards/alto/news.html#4-2-released

[6] See https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-sourceDoc.html

[7] Including when the images are distributed within the IIIF framework.

[8] Logical structure reconstruction can be performed semi-automatically (see the pipeline built for the LECTAUREP project called “LEPIDEMO”, https://github.com/lectaurep/lepidemo), or automatically with tools such as GROBID (https://github.com/kermitt2/grobid).

etc.) is a crucial step before disambiguation which further permits to build links with open general or domain-specific knowledge bases. These steps allow for later explorations of the text with data mining technologies.

Once all the layers of an edition are connected into the same TEI file, edited documents can be posted online with softwares like *TEI Publisher* (Turska et al., 2016; Chiffoleau et al., 2021). It provides a fully customizable environment where templates generate “views” based on the content of the XML files. With the aforementioned TEI structure, we propose an edition template containing:

- a flat representation of the transcription,
- an imitative representation of the transcription based on SVG[9] integrating the layout of the pages,
- a diplomatic edition of the source document, based on the content of the body element, and
- a facsimile, using the IIIF protocol (Fig. 3).

We would like to take the opportunity of presenting a short paper during the DH2022 international conference to subject our framework (Fig. 4) -and its robustness to different writing systems- to the scrutiny of the DH community. In particular, we believe that our proposition addresses challenges raised by Open Science, primarily the necessity to gain better control over every step within complex pipelines that involve various tools, thus facilitating reproducibility. A paradigm revolving around a pivotal element, like a TEI file grouping the different results, frees us from the constraint of a linear progression by maintaining multiple entry points in the workflow.

[9] An XML-based markup language, see the Scalable Vector Graphics (SVG) 2 recommandations at https://www.w3.org/TR/SVG2/ ; we wish to point at the fact that working with SVG when displaying transcriptions allows us to deal with different writing systems and languages.

Figures

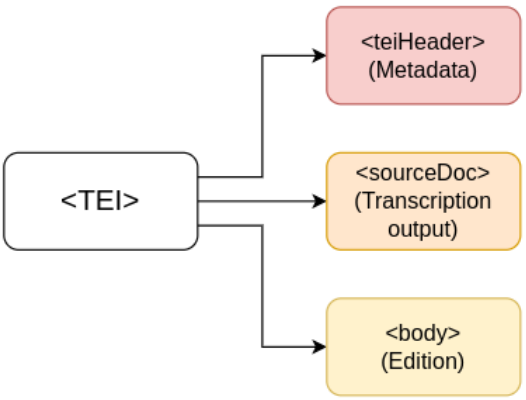

Fig. 1: TEI as a threefold structure.

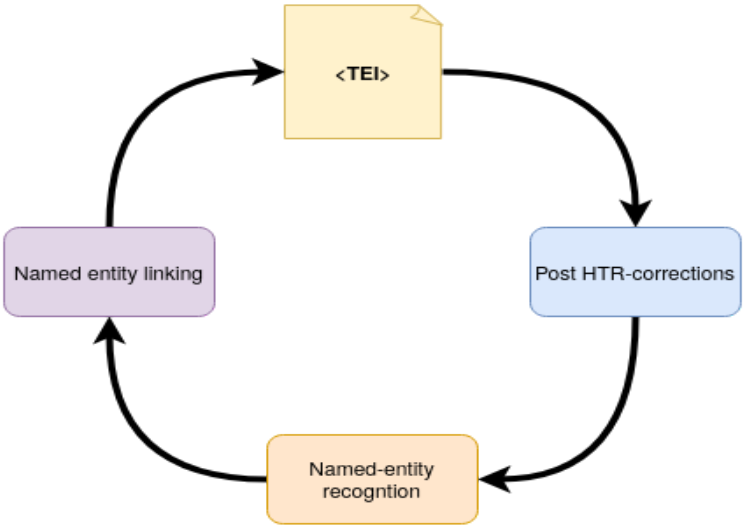

Fig. 2: Virtuous circle for the enriched TEI document.

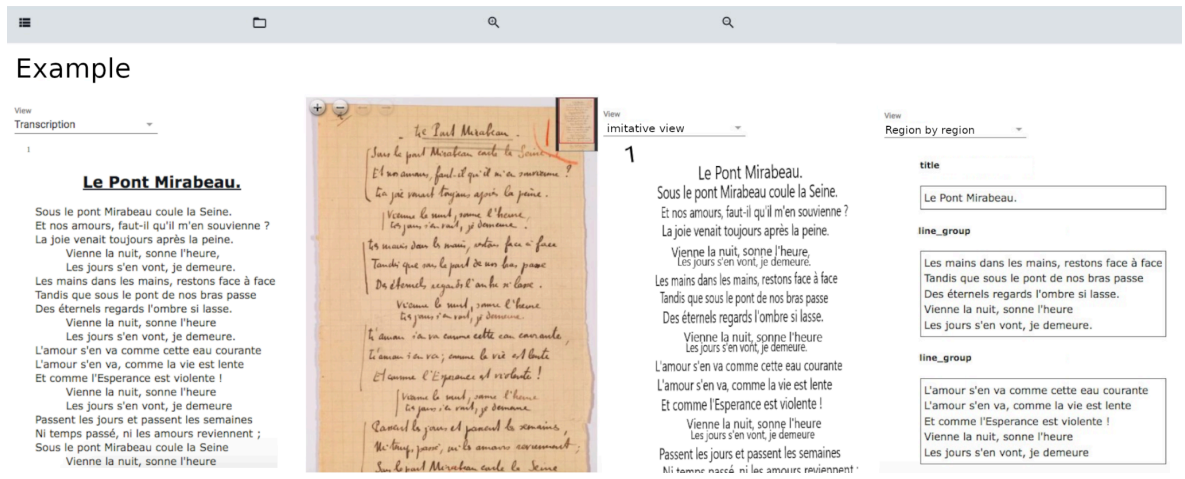

Fig. 3: A mock-up showing the four different views potentially available in an application like TEI-Publisher.

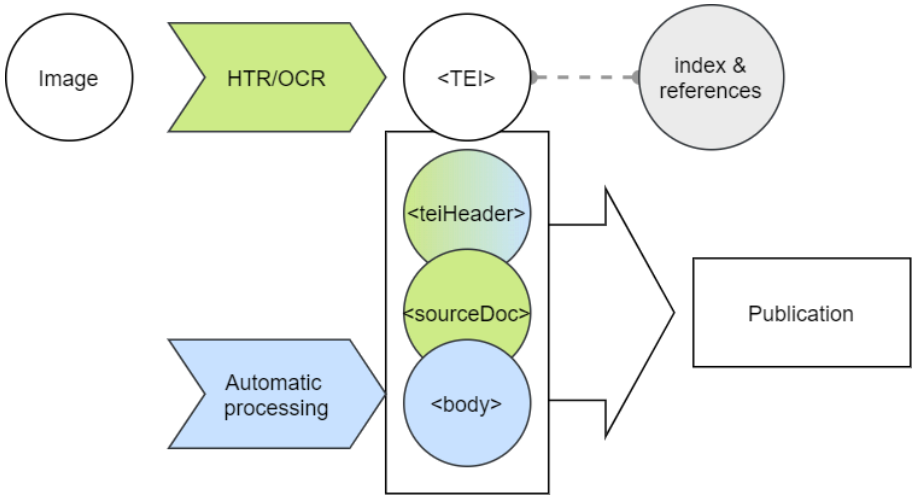

Fig. 4: Simplifying the workflow by using TEI from the beginning.

Cited tools

Chagué, A., & Scheithauer, H. (2021). *LEPIDEMO, a Pipeline Demonstrator for LECTAUREP to go from eScriptorium to TEI-Publisher* (1.0) [Jupyter Notebook]. https://doi.org/10.5072/zenodo.977657

e-editiones. (2021). *Eeditiones/tei-publisher-app* (7.1.0) [XQuery]. e-editiones.org. https://github.com/eeditiones/tei-publisher-app (Original work published 2020)

Kiessling, B. (2021). *Mittagessen/kraken* (3.0.6) [Python]. https://github.com/mittagessen/kraken (Original work published 2015)

Lopez, P. (2008). *GROBID* (0.7.0) [Java]. https://github.com/kermitt2/grobid

Terriel, L. (2021). *Semantic@* (1.0) [Python]. https://github.com/Lucaterre/semanticat (Original work published 2021)

Tissot, R. (2021). *Scripta/eScriptorium* (0.10.2a) [Python]. https://gitlab.com/scripta/escriptorium/-/tree/v0.10.2a

Bibliography

Balogh, D., & Griffiths, A. (2020). *DHARMA Encoding Guide for Diplomatic Editions* [Report, EFEO ; Humboldt-Universität (Berlin) ; CEAIS - Centre d'Études de l'Inde et de l'Asie du Sud]. https://halshs.archives-ouvertes.fr/halshs-02888186

Camps, J.-B. (2021). *Gallic(orpor)a : Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue*. Inauguration du BnF DataLab, Paris. https://www.academia.edu/58990010/Gallic_orpor_a_Extraction_annotation_et_diffusion_de_l_information_textuelle_et_visuelle_en_diachronie_longue

Carius, J.-C. (2020). Plateforme d'éditions enrichies à l'INHA : Premier point d'étape d'un projet en cours d'élaboration [Billet]. *Numérique et recherche en histoire de l'art*. https://numrha.hypotheses.org/1107

Causer, T., Grint, K., Sichani, A.-M., & Terras, M. (2018). « Making such bargain »' : Transcribe Bentham and the quality and cost-effectiveness of crowdsourced

transcription. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx064

Chiffoleau, F., Baillot, A., & Ovide, M. (2021). A TEI-based publication pipeline for historical egodocuments -the DAHN project. *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. https://hal.archives-ouvertes.fr/hal-03451421

Clanuwat, T., Lamb, A., & Kitamoto, A. (2019). KuroNet : Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning. *arXiv:1910.09433 [cs]*. http://arxiv.org/abs/1910.09433

Ehrmann, M., Romanello, M., Flückiger, A., & Clematide, S. (2020). Extended Overview of CLEF HIPE 2020 : Named Entity Processing on Historical Newspapers. *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*. 11th Conference and Labs of the Evaluation Forum (CLEF 2020), [online event]. https://doi.org/10.5281/ZENODO.4117566

Frontini, F., Brando, C., & Ganascia, J.-G. (2015). Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts. In A. Zucker, I. Draelants, C. F. Zucker, & A. Monnin (Éds.), *First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*. Arnaud Zucker and Isabelle Draelants and Catherine Faron Zucker and Alexandre Monnin. https://hal.archives-ouvertes.fr/hal-01203358

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017). Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, *04*, 19-24. https://doi.org/10.1109/ICDAR.2017.307

Pletschacher, S., & Antonacopoulos, A. (2010). *The PAGE (Page Analysis and Ground-truth Elements) format framework*. 257-260. https://doi.org/10.1109/ICPR.2010.72

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M., & Vidal, E. (2017). *ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset*. 1383-1388. https://doi.org/10.1109/ICDAR.2017.226

Scheithauer, H., Chagué, A., Gabay, S., Romary, L., Janes, J., & Jahan, C. (2021). From page to content – which TEI representation for HTR output? *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. https://hal.archives-ouvertes.fr/hal-03380807

Seaward, L. (2017). Project Update – teaching a computer to READ Bentham. *UCL Transcribe Bentham*. http://blogs.ucl.ac.uk/transcribe-bentham/2017/06/09/project-update-teaching-a-computer-to-read-bentham/

Stern, R. (2013). *Identification automatique d'entités pour l'enrichissement de contenus textuels* [Phdthesis, Université Paris-Diderot - Paris VII]. https://tel.archives-ouvertes.fr/tel-00939420

Stokes, P. A., Kiessling, B., Ezra, D. S. B., Tissot, R., & Gargem, E. H. (2021). The eScriptorium VRE for Manuscript Cultures. *Classics@ Journal*. https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/

Turska, M., Cummings, J., & Rahtz, S. (2016). Challenging the Myth of Presentation in Digital Editions. *Journal of the Text Encoding Initiative*, *Issue 9*, Article Issue 9. https://doi.org/10.4000/jtei.1453

Yin, F., Wang, Q.-F., Zhang, X.-Y., & Liu, C.-L. (2013). *ICDAR 2013 Chinese Handwriting Recognition Competition*. 1464-1470. https://doi.org/10.1109/ICDAR.2013.218