Keeping it open: a TEI-based publication pipeline for historical documents

Floriane Chiffoleau

Keeping it open: a TEI-based publication pipeline for historical documents

Floriane Chiffoleau

October 2021

Abstract:

Following the emergence of numerous projects to exploit historical archives, books or similar contents, as well as the exponential needs for digital tools tailored for those tasks, the DAHN project (Dispositif de soutien à l'Archivistique et aux Humanités Numériques) developed a complete open-source pipeline made of tools and methods making it possible to present a digital scholarly edition of scanned handwritten material. Composed of six steps (digitization, segmentation, transcription, post-OCR processing, encoding, and publication) and centered on historical documents, and more particularly on ego documents, this pipeline has been built around TEI, which works as a pivot format, to ensure its robustness, sustainability, and reusability. More than just encoding in TEI, we also choose tools compatible with it, such as eScriptorium for segmentation/transcription or TEI Publisher for the publication. To further help the people working with the pipeline, we also heavily documented the development of the pipeline, as well as its steps, to ease its reuse.

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

Biography:

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 before starting 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).

Email: floriane.chiffoleau@inria.fr

Over the past years, the need for digital tools tailored for Humanities research has grown dramatically. In order to exploit their archives, organize their collections, present editions that fulfill the requirements of each of the concerned disciplines as well as reusable material, the need for easy-to-use tools for non-programmers to digitize, transcribe and publish their corpus has given way to different initiatives aiming at tackling this challenge (Pierazzo, 2019).

However, the multiplication of tools for the same task (such as several OCR systems (Assefi 2016)), the difficulty to find the right one for a specific task and the privatization of some of these tools raised obstacles for researchers trying to settle for suitable technical solutions and reach their goals.

To respond to that demand, we propose a pipeline for digital scholarly editing of historical documents, created in the context of the DAHN project, a technological and scientific collaboration between Inria, Le Mans Université and EHESS, funded by the French Ministry of Higher Education, Research and Innovation. The project aims at facilitating the digitization of data extracted from archival collections and their dissemination to the public in the form of digital documents in various formats and as an online edition. In order to meet these expectations, the project team developed an open-source TEI-based workflow[1] with easy-to-use, freely accessible tools and methods, and a thorough documentation. The workflow that was developed this way (fig. 1) is divided into six steps, each relying on a dedicated process in order to offer a determined path to those who need it:

- digitization, with IIIF servers (our go-to is NAKALA)
- segmentation and transcription, with Kraken and its interface eScriptorium
- post-OCR processing, with Python and (currently) the *pyspellchecker* library
- encoding, in TEI XML, using Python, regular expressions and spreadsheets
- publication, with TEI Publisher

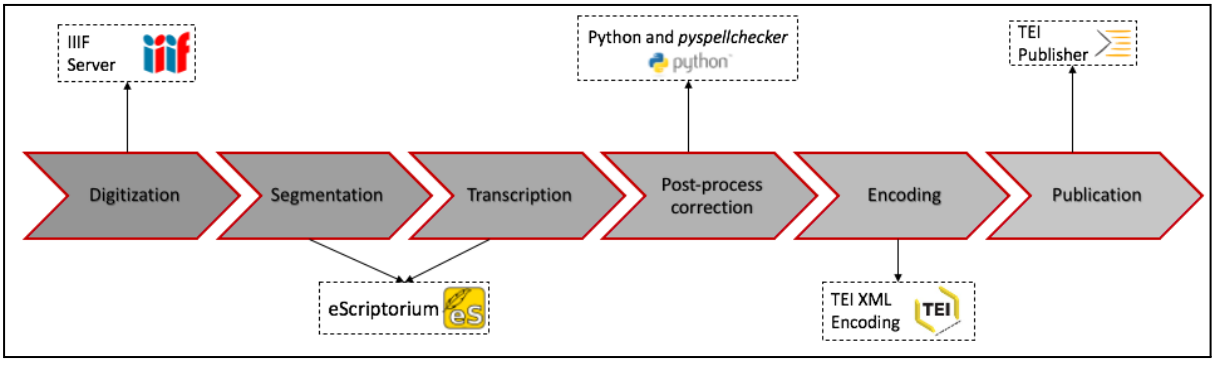


Figure 1 : Pipeline for the digital edition of historical documents

[1] https://github.com/DiScholEd/pipeline-digital-scholarly-editions

The pipeline was developed based on the guidelines and principles of the *Text Encoding Initiative (TEI)*. Using a standard for the representation of texts contributes to moving swiftly from one step to the next without any information loss (Scheithauer 2022). It also ensures better sustainability for the corpus. We also decided to train this pipeline around ego documents, which are texts written in the first person, from the private sphere, that can be easier to process, thanks to the specificities they can contain in form and content (e.g., structure, repetitive parts). Finally, extensive documentation is a valuable asset for researchers and developers to reuse and/or adapt elements of the pipeline: the development of the pipeline (from its onset) as well as its usage rules have been thoroughly documented and are easily accessible[2]. Additionally, all the project code and resources have been deposited in a GitHub organization dedicated to the project.[3]

Therefore, the goal of this paper is to go through the steps taken to build this TEI-based publication pipeline. In the following, we will see that after experimenting with the pipeline, we produced our transcriptions using the eScriptorium interface. Following that, we worked on our encoding through TEI-XML and Python scripts (with libraries such as BeautifulSoup or pyspellchecker) for the automation of the process to finally publish those corpora, helped by *prêt-à-porter* applications for TEI files (TEI Publisher).

1. Pipeline experiments: the corpora at hand

In order to be able to characterize the specificities and impact of ego documents in the setting up of our document processing pipeline, we chose to rely on two complementary text collections. First, we selected a homogeneous source (Paul d'Estournelles de Constant's letters to Nicholas Murray Butler[4]) which would require a complete recognition/structuring/publication process, with a medium-level complexity in terms of HTR, namely in the form of scans of typewritten texts. In addition, with the idea in mind to have a reference encoding at our disposal, we selected an existing digital scholarly edition (Berlin Intellectuals[5]).

The first corpus we relied upon for the creation of the pipeline (which we consider as our baseline) is a correspondence dating back to the beginning of the twentieth century, and more precisely World War I. The Paul d'Estournelles de Constant corpus (Tison 2015, Tison and Akhund 2018), named after the main author of the presented correspondence, is a long

---

[2] The Digital Intellectuals blog, category DAHN: https://digitalintellectuals.hypotheses.org/category/dahn/work-progress
[3] https://github.com/DiScholEd
[4] https://github.com/DiScholEd/paul-destournelles-de-constant
[5] https://github.com/DiScholEd/berlin-intellectuals

and lasting correspondence. It spans from August 1914 to May 1924 and includes about 1500 letters[6]. The main topics discussed in this correspondence are political; it deals with the war and its aftermath, current national and international affairs, diplomatic missions as well as issues of everyday life. The 1500 letters were all sent by Nobel Peace Prize laureate Paul d'Estournelles de Constant to his friend, Nicholas Murray Butler[7]. This corpus was a gold mine for our experiment, since, at the beginning of our project, the only available information was photographs of the letters together with a few transcriptions, both supplied by historian and project colleague Stéphane Tison. Based on this scarce material, we had the opportunity to test the pipeline from segmentation to publication while professional scanning was carried out by the holding institution and made available for publication[8]. While working on this corpus, we adjusted the segmentation model, developed a transcription model suited to the writing style of the corpus (typewritten documents), and created generic scripts to correct transcription and encoding. Finally, once the documents were encoded, we built our publication application in such a manner that it would fittingly display the content of the encoded documents and the type of information that we intended to make available for the online publication.

Another corpus was also brought into the development of the application to allow us to further test various stages of our pipeline. The edition “Letters and Texts. Intellectual Berlin around 1800” had already been online for several years and was lacking a sustainable hosting and development environment, despite being fully encoded and published (Baillot, since 2012). This edition is composed of correspondence as well as other types of handwritten documents written by people from intellectual milieus of the first half of the nineteenth century. It posed a challenge in the sense that it was not limited to ego documents, and required the developed solution to be able to process and display different types of documents. We updated the encoding to the latest TEI standard, and used it for the digitization part of the process: the information was well-structured and helped gather metadata needed to add the images of the corpus to the IIIF server (NAKALA) efficiently[9].

Using these two main corpora facilitated the elaboration of our pipeline. Once they were integrated and/or updated, we were able to measure how well it works thanks to new

---

[6] This is not the only correspondence held by d’Estournelles during his life, as he wrote regularly to other people, as well as to Murray Butler before this period of time, but this corpus contains the collection of numbered and classified letters created by d’Estournelles himself for the time and called “*La guerre 1914-1918. Collection de mes lettres numérotées au président Butler*“

[7] Murray Butler did respond to those letters, but these responses are kept in a different institution and are not part of our corpus (Some can still be read in (Tison and Akhund 2018))

[8] Archives départementales de la Sarthe, Sous-série 12J: http://archives.sarthe.fr/

[9] The scripts used to do that can be found here, and the documentation detailing the process here.

projects, more explicitly presented below, and that were interested mostly in the last phase of the process. The perks of a TEI-based publication pipeline were quickly noticeable, because the new projects were displayed, without any major issues, on our application. It only needed some TEI adaptations to match our encoding. Some of the new projects decided to develop their own publication application, as their files were not strictly ego documents, or they had a strong interest for features we did not implement, as was the case for TIME US[10] and LECTAUREP[11]. Other projects decided to choose our pipeline and platform as a host and simply implemented minor modifications in their TEI-encoding in order to be more compliant with the features we developed, as is the case with the EHRI corpus[12], the Rochlitz manuscript[13] and My War by Charles Bruneau[14].

2. Obtaining digital texts: the eScriptorium interface

We concentrate here on the definition of a reliable process to move from digital images to digital texts. Taking into account the specificities of the corpora used in our document processing pipeline, we used an open-source HTR interface, eScriptorium[15] that works for both printed and handwritten texts and offers practical outputs for subsequent uses.

Over the past decades, numerous historical documents from museums, archives, and libraries were digitized in order to increase their accessibility for the public. To foster work based on these documents, tools to transcribe them have been developed, be it OCR (printed documents) or HTR (handwritten documents) (Romero et al. 2011, Nockels 2022). Several OCR systems exist that are widely used for printed documents because they are accessible and easy to use, such as Tesseract or ABBYY FineReader (Heliński 2012; Assefi 2016) for instance. However, we identified that an OCR system would not fit the needs we had and therefore decided for a more elaborate HTR system. The main requirements were openness to as many different projects as possible, and not only to projects with corpora made of printed documents. Hence, finding a suitable HTR system was key, and we were testing the following two: Transkribus and Kraken/eScriptorium. Transkribus offers segmentation, text recognition and multiple exports, as eScriptorium does too, as well as some other options like Keyword Spotting (KWS) (Kahle et al. 2017). However, it was ultimately not chosen for our pipeline as we preferred eScriptorium first because it is fully

[10] The Time-Us blog: https://timeus.hypotheses.org
[11] The Lectaurep blog: https://lectaurep.hypotheses.org
[12] https://discholed.huma-num.fr/exist/apps/discholed/index.html?collection=ehri
[13] https://discholed.huma-num.fr/exist/apps/discholed/index.html?collection=rochlitz
[14] https://discholed.huma-num.fr/exist/apps/discholed/index.html?collection=maguerre
[15] The eScriptorium interface: https://escriptorium.inria.fr/

free of charge, and also because the interface is partly developed by members of ALMAnaCH.

eScriptorium is an open-source web interface for collaborative and automatic transcription projects, "developed in the frame of the Scripta PSL program at PSL University" (Kiessling et al., 2019) and based on the OCR software *Kraken*[16] (Kiessling, 2019). This web interface provides an adequate alternative to users who want to segment and transcribe their corpus, but are not comfortable with a command line interface. On top of those basic functions for the text recognition of corpora, *Kraken* and the eScriptorium interface also offer the benefit of being able to create ground truth and to train models for specific corpus (Fischer, 2010), which is perfect for Handwritten Text Recognition (HTR). Thus, with this possibility to train models and to work with HTR or OCR, this part of the pipeline can be open to more projects.

The export formats offered by eScriptorium also resonate with our TEI-based publication pipeline because many are compatible with TEI. Firstly, there is the possibility of a simple text output. The encoding will start from scratch, using text structure and regular expression, like it was the case for the encoding of Paul d'Estournelles de Constant' corpus.[17] Secondly, eScriptorium also works with XML outputs, namely ALTO[18] and PAGE[19] that are formats that register regions, lines, baselines and encoded texts, through established and regulated tags. Helped by an XSL-Python transformation pipeline created by members of the ALMAnaCH team[20] (Chagué & Scheithauer, 2021) and with the cooperation of Manon Ovide (an intern of the DAHN project), those XML outputs can be transformed according to the TEI Guidelines. The output TEI XML file will contain a teiHeader and a text element, as per usual, but a sourceDoc element[21] will also be added, containing the coordinates on the facsimile and the corresponding text, provided thanks to the regions and lines established during the segmentation.

3. Completely and accurately rendering the transcription: encoding with TEI XML

In order to be able to exploit at full potential a transcribed corpus, it is advisable to encode it in TEI XML, for better readability and enhanced access. As this task can be tedious, we propose with our document processing pipeline some options to speed up this work, by using regular expressions, metadata information and thorough guidelines, all done while following at every step the TEI guidelines.

---

[16] The Kraken Website: http://kraken.re
[17] https://digitalintellectuals.hypotheses.org/3891
[18] A description of text OCR and layout information of pages for digitized material (Source: Wikipedia)
[19] An XML-based page image representation framework that records information on image characteristics in addition to layout structure and page content (Source: Pletschacher & Antonacopoulos 2010)
[20] Page2tei: https://github.com/TEI4HTR/page2tei
[21] https://tei-c.org/Vault/P5/4.3.0/doc/tei-p5-doc/en/html/ref-sourceDoc.html

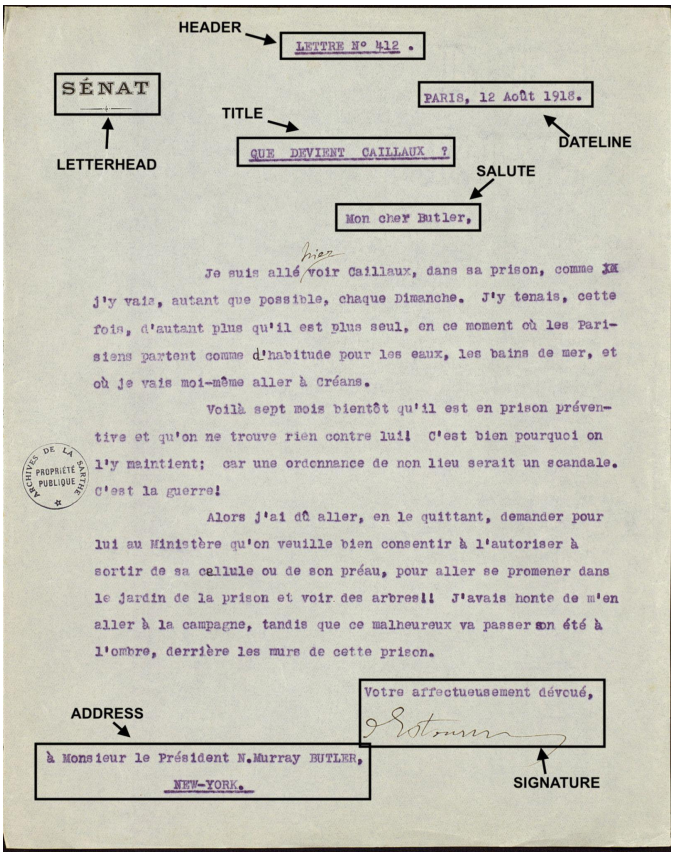
We chose TEI because of its status as the standard for the representation of digital texts and for its thoroughly documented guidelines (TEI Consortium, 2021). Furthermore, on top of those guidelines and in order to clarify the rules of encoding of the ego documents, the recently TEI-recommended type of files we are working on (Stadler 2016), we also developed an encoding guide for ego documents. It concentrates, in an ODD, all the tags that can be used, in the header or in the body, to encode a corpus of ego documents, as well as suggestions of a specific tag in some cases where the TEI Guidelines provide several encoding options. In that way, when a project working with ego documents tries to use our pipeline, from start to finish, it would not have to lose valuable time by searching in all the guidelines to guess how to encode what (Durusau, 2006) and it will just need to follow the rules we stated, which will also be easier if the members of the project decide to use our publication pipeline, as it will guarantee an adequate display of their files.

Moreover, to automate part of the encoding process, we wrote Python scripts that encode part of the files by using the recurrence and regularity of its content. For example, in the same corpus, most of the information from the header should not differ from one file to another because it usually has the same origin, the same people working on it, the same structure, etc. It is then possible, with a script and a spreadsheet containing the basic information from the corpus (number of the letter, date in multiple formats, number of pages, etc.), to generate the corpus header content all at once. In the script, the generic information of the corpus will have been added (conservation place, rules of encoding, project information, etc.) and thanks to the spreadsheet, new information are added. For example, if we consider what is presented below, the content of “Letter” and “FullDate” will be used for the naming of the XML file; “French Date” will be used for the French title of the file (<title xml:lang=”fr”>) and “English Date” for the English title; “Date” will be used in the date tag such as the one in the <correspDesc>; “Pages” will be used in a tag mentioning how many pages in the encoded document. However, even though the script fills automatically many parts of the <teiHeader> element, it does not fill every field and some parts of the metadata will have to be encoded manually.

	Letter	Date	French Date	English Date	Full Date	Pages
1	1	1914-08-15	15 août 1914	August 15, 1914	15août1914	12
2	2	1914-09-03	3 septembre 1914	September 3, 1914	3septembre1914	6
3	3	1914-09-08	8 septembre 1914	September 8, 1914	8septembre1914	6
4	4	1914-09-11	11 septembre 1914	September 11, 1914	11septembre1914	9
5	5	1914-09-18	18 septembre 1914	September 18, 1914	18septembre1914	9
6	6	1914-09-24	24 septembre 1914	September 24, 1914	24septembre1914	9
7	7	1914-10-19	19 octobre 1914	October 19, 1914	19octobre1914	16
8	8	1914-11-02	2 novembre 1914	November 2, 1914	2novembre1914	9
9	9	1914-11-13	13 novembre 1914	November 13, 1914	13novembre1914	5
10	9bis	1914-11-16	16 novembre 1914	November 16, 1914	16novembre1914	7

Figure 2: Spreadsheet of the information for the header metadata

For the body, our work has been considerably helped by the uniformity of the structure of the corpus on which we developed our workflow (Paul d'Estournelles de Constant), — as it kept the same structure for the ten years of writing, with a header, a letterhead, a dateline, a title, a salute, text and then address, signature, and again the address (fig. 3) —, as well as by the work of the TEI Special Interest Group (SIG) Correspondence[22] and their manual "Encoding correspondence"[23].



[22] https://tei-c.org/Activities/SIG/Correspondence/
[23] https://encoding-correspondence.bbaw.de/v1/index.html

Figure 3: First page of the letter 412 from Paul d'Estournelles' correspondence (the repetitive parts have been framed)

With the help of some regular expressions, writing an encoding script was not too hard, as it enables us to encode 90% of the corpus expeditiously.[24] Once again, this is also documented and explained, step by step, so that anyone with ego documents with repetitive sequences such as ours, using our pipeline, can write their own regular expressions and encode their texts. Our goal is that the script could be otherwise useful for people with corpus that are not ego documents, as it provides at least basic encoding for texts by using the punctuation.

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

As previously stated, developing a corpus following the guidelines of the *Text Encoding Initiative* is a wise choice as much for its role as a standard for the representation of digital texts as for the opportunity it embodies for collaborative and remote work. Despite it, this TEI encoding is pointless without the possibility to make it available and public for the community (Turska, Cummings & Rahtz 2016). Furthermore, in the current landscape, the use of CMS, that could be considered *prêt-à-porter* applications, become more prominent (Pierazzo, 2019), as it can make the process of publication faster and easier, and this is where TEI Publisher[25] comes into play in our workflow.

TEI Publisher is a tool developed for scholars and editors to offer them the possibility to publish their work, without having to become a programmer or to hire one to develop an entire website[26]. For a digital humanist, TEI Publisher offers an interesting compromise between developing a website for the corpora of the pipeline without any extended programming skills, while still being able to add some features and customization to offer a unique product for the pipeline.

Moreover, TEI Publisher is a tool made with and for TEI. It is based on the TEI processing model, which is "used to document the processing model intended for a particular element"[27]. It is composed of many elements that work as conditions or supplements for the element (fig. 4).

---

[24] Those scripts are available in the repository of our project: https://github.com/DiScholEd/pipeline-digital-scholarly-editions/tree/master/encoding/scripts

[25] The TEI Publisher: https://teipublisher.com/index.html toolbox: https://teipublisher.com/exist/apps/tei-publisher/index.html

[26] The TEI Publisher toolbox: https://teipublisher.com/exist/apps/tei-publisher/index.html

[27] The TEI Processing model: https://tei-c.org/Vault/P5/4.3.0/doc/tei-p5-doc/en/html/TD.html#TDPMPM

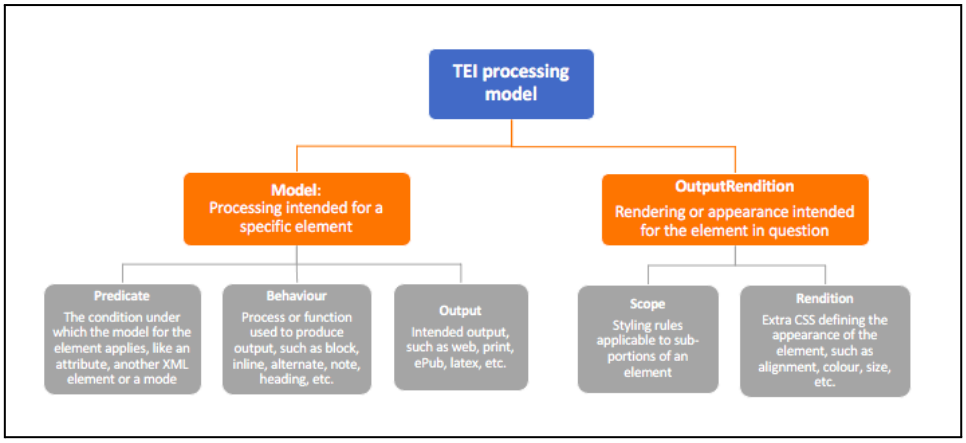


Figure 4: The TEI Processing model and its components

The presence of the attribute "predicate" offers the possibility to propose multiple displays for a single element, according to its position in the XML tree, its relationships to other elements or its attributes. This can even be reinforced by the existence of 'modes' in TEI Publisher. Those modes are directly linked to the templates used to display the XML files, and they represent extra parameters for an element. With the modes, an element can be displayed in multiple forms. For example, as shown below, in the case of the element “del” (deletion)[28], we can specify that when it has the attribute @rend="strikethrough", the content of the element will appear crossed for the mode "diplomatic version (corrections)". In the other cases, with the mode "reading version (original/entities)", we choose to not display it at all.

[28] https://tei-c.org/Vault/P5/4.3.0/doc/tei-p5-doc/en/html/ref-del.html

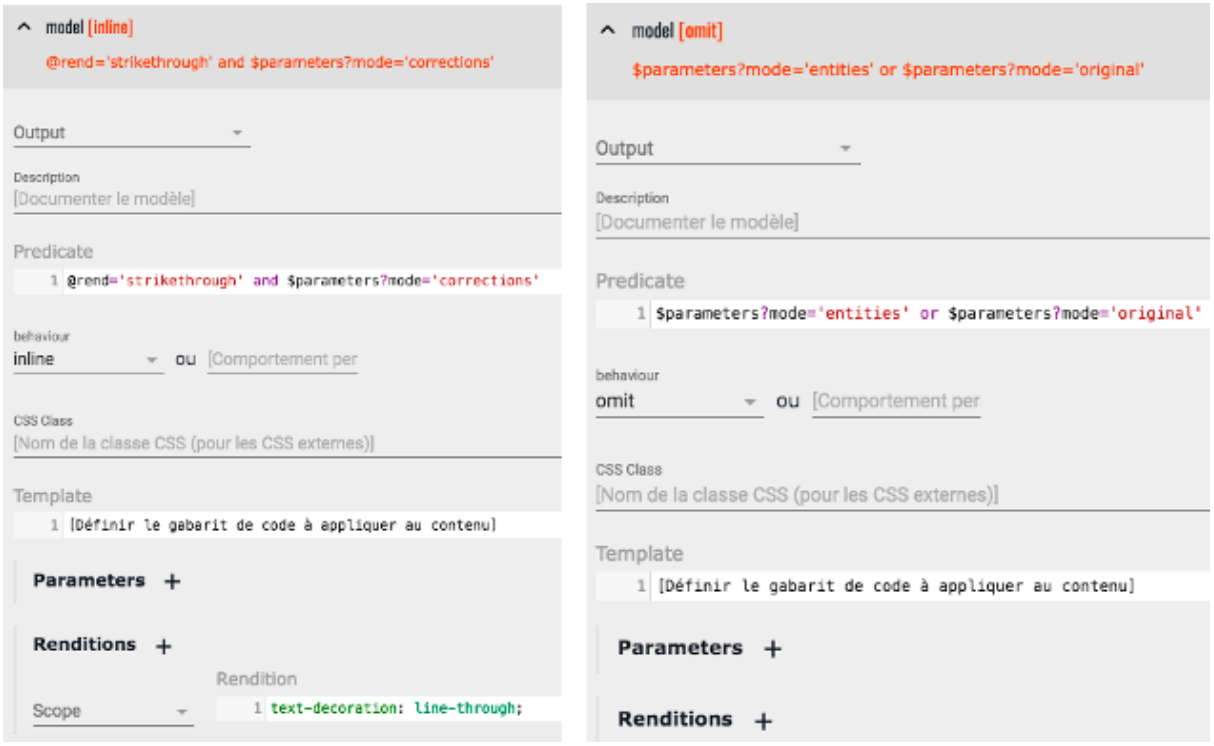


Figure 5: TEI processing models for the element “del”

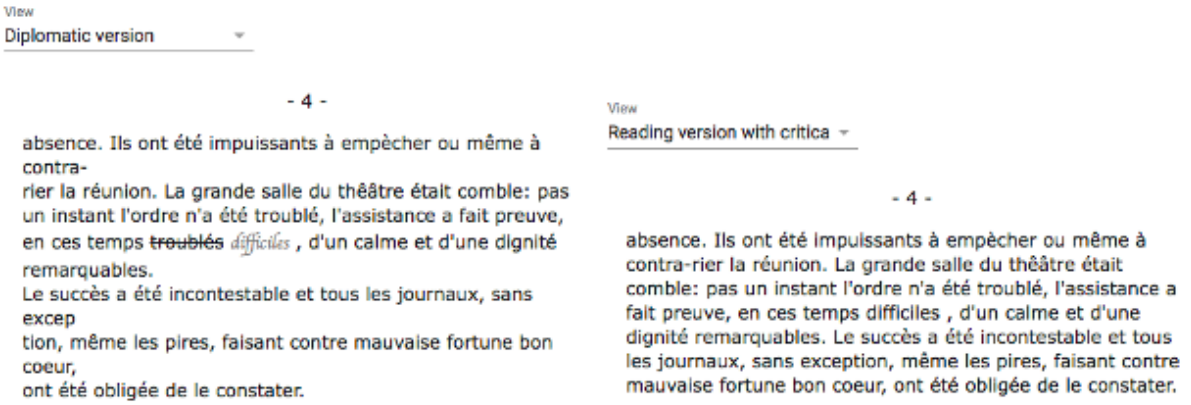


Figure 6: Display of the element “del” in the diplomatic (strikethrough) and reading (omitted) version

The attribute 'behaviour' is also an important part of this declaration because it can determine how we want the element to be interpreted. The most common display would be 'inline', to ordinarily integrate the element in its XML tree parent. We can also have it appear as a unique entity with 'block', as a title with 'heading', as a link with 'link' (provided we have a URI) or even not at all, with 'omit' (as seen in fig. 5).

Among the other attributes attached to 'model', the output gives us the choice to add the instructions for the web, for one of the downloadable outputs of the file or both. This enables

us to not only think about the user experience on the website, but also about what they can access and retrieve from our website and in what ways.

Finally, the TEI processing model also proposes 'output rendition' which gives us the opportunity to add size, color, emphasis, boldness, alignment, underlining and other kinds of effects to the element. We then either have it match the structure and presentation of the facsimile (and render what was defined in some attributes of the encoding) or focus on specific elements (for example, all the named entities appearing in bold for a view focused on it).

This TEI processing model is the reason behind our selection of TEI Publisher for our publication pipeline, and it turned out to be an informed choice. The richness of information offered by the TEI XML encoding is only useful if all of those specifications (and more) can be interpreted, all at once, and this is what we obtain here. With predicate, behavior, modes and rendition, we have the possibility to render the text in many ways that can be parts of multiple views, instead of confusing the user with all the elements mixed in a single view.

Conclusion: a minimalist pipeline for a maximum efficacy

As previously demonstrated, our pipeline for the digital edition of historical manuscripts relies on defined tools to perform correctly, some of them however being more oriented towards working with ego documents.

Moreover, its heavy ties with the *Text Encoding Initiative*, combined with the documentation and the encoding schema we created[29], ensure consistency and sustainability. As the TEI standard has developed over time, every little modification made was recorded, notified and documented. In this way, if the documentation or one of the transformation scripts needs to be adjusted to the new TEI Guidelines, it will be easier to carry on that change. With this pipeline and the advantages of the TEI, we want to offer to researchers the possibility to have linked tools at their disposal, provided with documentation and examples, so that they can more easily exploit the corpus they are working on. For the more specific case of ego documents, our goal is to go even further by proposing a space for publication, with our TEI Publisher application, provided they followed our guidelines, as those steps are fully linked, from the ODD for the encoding to the ODD for the publication display.

[29] https://github.com/DiScholEd/pipeline-digital-scholarly-editions/tree/master/encoding/guidelines

References:

Assefi, Mehdi. 2016. OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. ISCV.

Baillot, Anne. ed., Letters and texts. Intellectual Berlin around 1800. www.berliner-intellektuelle.eu, last online consultation: 2022-03-31

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

Durusau, Patrick. 2006. “Why and How to Document Your Markup Choices.” In *Electronic Scholarly Editing*, edited by Lou Burnard, Katherine O’Brien O’Keeffe, and John Unsworth, 299–309. New York: MLA.

Fischer, Andreas, Emanuel Indermühle, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. 2010. Ground truth creation for handwriting recognition in historical documents. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems* (*DAS '10*). Association for Computing Machinery, New York, NY, USA, 3–10. DOI:https://doi.org/10.1145/1815330.1815331

Heliński, Marcin, Milosz Kmieciak and Tomasz Parkola. “*Report on the comparison of Tesseract and ABBYY FineReader OCR engines.*” (2012). https://api.semanticscholar.org/CorpusID:6341954

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

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

Kiessling Benjamin. 2019. Kraken-an universal text recognizer for the humanities. In ADHO, Éd., Actes de Digital Humanities Conference 2019 - DH2019, Utrecht, Pays-Bas.

Nockels, J., Gooding, P., Ames, S. *et al.* 2022. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Arch Sci* 22, 367–392. https://doi.org/10.1007/s10502-022-09397-0

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⟩

Pletschacher, Stefan, and Apostolos Antonacopoulos. 2010. “The PAGE (Page Anal-ysis and Ground-Truth Elements) Format Framework”. In: *2010 20th Interna-tional Conference on Pattern Recognition*, pp. 257–260. DOI:10.1109/ICPR.2010.72

Romero, Verónica, Nicolás Serrano, Alejandro H. Toselli, Joan Andreu Sánchez, and Enrique Vidal. 2011. Handwritten Text Recognition for Historical Documents. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 90–96, Hissar, Bulgaria. Association for Computational Linguistics.

Scheithauer, Hugo, Alix Chagué and Laurent Romary. 2022. Which TEI representation for the output of automatic transcriptions and their metadata? An illustrated proposition. ⟨hal-04001303v2⟩

Stadler, Peter, Marcel Illetschko and Sabine Seifert. “Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>”, *Journal of the Text Encoding Initiative* [Online], Issue 9 | September 2016 - December 2017, Online since 24 September 2016, connection on 15 February 2022. URL: http://journals.openedition.org/jtei/1433; DOI: https://doi.org/10.4000/jtei.1433

TEI Consortium 2021. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.3.0. Last updated 31st August 2021. N.p.: TEI Consortium. https://tei-c.org/Vault/P5/4.3.0/doc/tei-p5-doc/en/html/.

Tison, Stéphane, and Nadine Akhund. 2018. *En guerre pour la paix. Correspondance Paul d’Estournelles de Constant et Nicholas Murray-Butler 1914-1919*. Alma Editeur. URL:https://hal.archives-ouvertes.fr/hal-01865430.

Tison, Stéphane. Paul d’Estournelles de Constant. 2015. 1914, un pacifiste face à la guerre. Stéphane Tison (dir.). *Paul d’Estournelles de Constant, Prix Nobel de la paix 1909. Concilier les nations pour empêcher la guerre (1878-1924)*, Presses universitaires de Rennes, pp.133-161. ⟨hal-02275381⟩

Turska, Magdalena, James Cummings and Sebastian Rahtz. “Challenging the Myth of Presentation in Digital Editions”, Journal of the Text Encoding Initiative [Online], Issue 9 | September 2016 - December 2017, Online since 24 September 2016, connection on 03 March 2022. URL: http://journals.openedition.org/jtei/1453; DOI: https://doi.org/10.4000/jtei.1453