From page to content - which TEI representation for HTR output?

Alix Chagué[1], Simon Gabay[2], Laurent Romary[1], Juliette Janes[3], Hugo Scheithauer[3,1], and Claire Jahan[3]

[1]INRIA, team ALMAnaCH (France)
[2]Université de Genève (Switzerland)
[3]PSL - École nationale des chartes (France)

We can anticipate that one of the essential sources for digital texts in the forthcoming future will result from the recent outstanding improvements of Handwritten Text Recognition (HTR) techniques as reflected with online services such as Transkribus (Kahle et al. 2017) or open source platforms such as eScriptorium (Kiessling et al. 2019) / Kraken (Kiessling 2019). Still, most of the existing HTR platforms rely, with good reasons, on layout oriented exchange formats such as Page XML[1] or ALTO[2] to express recognition outputs integrating both plain text content and layout information. From a TEI community point of view, it is thus important that we identify some precise guidelines for representing this content in TEI with a scenario in mind where we would want to be able to reuse a shared low-level (*i.e.* shallow text semantic) encoding across projects that may bear appropriate visualisation capacities or even be a common basis for more elaborate encodings according to the type of documents each project is dealing with. The present paper will show the results of confronting two digitisation projects that have relied upon the eScriptorium / Kraken platform and their attempts to map Page XML and ALTO content onto a stable TEI structure and address the following issues:

- integrating the proper metadata in the `<teiHeader>` (basic reference to the HTR software, trained models for the layout analysis and/or the transcription);
- mapping the layout format in the `<sourceDoc>` element, keeping track of all segments identified by the HTR process together with the corresponding text;
- experimenting with a controlled vocabulary for the description of the layout, as provided by the SegmOnto working group (Gabay et al. 2021);

[1]http://www.primaresearch.org/tools/PAGELibraries.
[2]https://www.loc.gov/standards/alto.

- keeping the link with external IIIF image servers;
- linking interpreted content in the `<text>` elements with the various components available in `<sourceDoc>`.

The paper will come with updated github resources (example files, XSLT transforms) at the time of the conference.

Data+code

Scripts and data are available at the following adresses:

- https://github.com/e-ditiones/Annotator.
- https://github.com/lectaurep/page2tei.
- https://github.com/lectaurep/lepidemo.

References

Gabay, S. et al. (Sept. 2021). "SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)". In: *Proceedings of the 1st International Workshop on Computational Paleography, IWCP@ICDAR 2021*. 1st International Workshop on Computational Paleography IWCP. Lecture Notes in Computer Science. Lausanne (Switzerland): Springer.

Kahle, P. et al. (2017). "Transkribus - a service platform for transcription, recognition and retrieval of historical documents". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 4. IEEE, pp. 19–24. URL: https://ieeexplore.ieee.org/document/8270253.

Kiessling, B. (July 2019). "Kraken - an Universal Text Recognizer for the Humanities". In: Utrecht, The Netherlands: Alliance of Digital Humanities Organizations (ADHO). URL: https://dev.clariah.nl/files/dh2019/boa/0673.html.

Kiessling, B. et al. (Sept. 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.