Between automatic and manual encoding: towards a generic TEI model for historical prints and manuscripts

Ariane Pinche[1], Kelly Christensen[2], and Simon Gabay[2]

[1]Ecole nationale des chartes | PSL (France)
[2]Inria (France)
[2]Université de Genève (Switzerland)

keywords : TEI, text extraction, linguistic annotation, digital edition, mass digitisation

Abstract

Cultural heritage institutions today aim to digitise their collections of prints and manuscripts (Bermès 2020) and are generating more and more digital images (Gray 2009). To enrich these images, many institutions work with standardised formats such as IIIF, preserving as much of the source's information as possible. To take full advantage of textual documents, an image alone is not enough. Thanks to automatic text recognition technology, it is now possible to extract images' content on a large scale. The TEI seems to provide the perfect format to capture both an image's formal and textual data (Janès et al. 2021). However, this poses a problem. To ensure compatibility with a range of use cases, TEI XML files must guarantee IIIF or RDF exports and therefore must be based on strict data structures that can be automated. But a rigid structure contradicts the basic principles of philology, which require maximum flexibility to cope with various situations.

The solution proposed by the *Gallic(orpor)a* project[1] attempted to deal with such a contradiction, focusing on French historical documents produced between the 15th and the 18th c. It aims to enrich the digital facsimiles distributed by the French National Library (BnF)[2] in two different ways:

- text extraction, including the segmentation of the image (layout analysis) with *SegmOnto* (Gabay, Camps, et al. 2021) and the recognition of the text (Handwritten Text Recognition) augmenting already existing models such as Pinche and Clérice (2021);
- linguistic annotation, including lemmatisation, POS tagging (Gabay, Clérice, et al. 2020), named entity recognition and linguistic normalisation (Bawden et al. 2022).

[1]`https://gallicorpora.github.io`.
[2]`https://gallica.bnf.fr`.

Our TEI document modelling has two strictly coercive automatically generated data blocks:

- the `<sourceDoc>` with information from the digital facsimile, which computer vision, HTR and segmentation tools produce thanks to machine learning (Scheithauer et al. 2021);
- the `<standOff>` (Bartz et al. 2021a) with linguistic information produced by natural language processing tools (Gabay, Suarez, et al. 2022) to make it easier to search the corpus (Bartz et al. 2021b).

Two other elements are added that can be customised according to researchers' specific needs:

- a pre-filled `<teiHeader>` with basic bibliographic metadata automatically retrieved from (i) the digital facsimile's IIIF Image API and (ii) the BnF's Search/Retrieve via URL (SRU) API.[3] The `<teiHeader>` can be enriched with additional data, as long as it respects a strict minimum encoding;
- a pre-editorialised `<body>` (fig. 1). It is the only element totally free regarding encoding choices.

```
<body>
  <div>
    <pb corresp="#page5"/>
    <note corresp="#page5_zone2" type="MarginTextZone">
      <lb corresp="#page5_zone2_line1"/>79/4120
    </note>
    <pb corresp="#page6"/>
    <ab corresp="#page6_zone1" type="MainZone">
      <hi rend="HeadingLine">
        <lb corresp="#page6_zone1_line1"/>BRADAMANTE,
        <lb corresp="#page6_zone1_line2"/>TRAGECOMDEDIE.
      </hi>
    </ab>
    <pb corresp="#page9"/>
    <fw corresp="#page9_zone1" type="RunningTitleZone">
      <lb corresp="#page9_zone1_line1"/>AV ROY.
    </fw>
    <ab corresp="#page9_zone2" type="MainZone">
      <lb corresp="#page9_zone2_line1"/>uiuront nostre siecle, les admira¬
      <lb corresp="#page9_zone2_line2"/>bles effets de vos heroiques ver¬
      <gap reason ="sampling"/>
    </ab>
  </div>
</body>
```

Figure 1: Example of a pre-editorialised `<body>`: Robert Garnier, *Tragédies*, Paris: Robert Estienne, 1582 [ark:/12148/bpt6k990549b].

By restricting certain elements and allowing others to be customisable, our TEI model can efficiently pivot toward other export formats, including RDF and IIIF. Furthermore,

[3] `https://catalogue.bnf.fr/api`

the <sourceDoc> element's strict and thorough encoding of all of the document's graphical information allows the TEI document to be converted into PAGE XML and ALTO XML files, which can then be used to train OCR, HTR, and segmentation models. Thus, not only does our TEI model's strict encoding avoid limiting philological choices, thanks to the <body>, it also allows us to pre-editorialise the <body> via the content of the <sourceDoc> and, in a near future, the <standOff>.

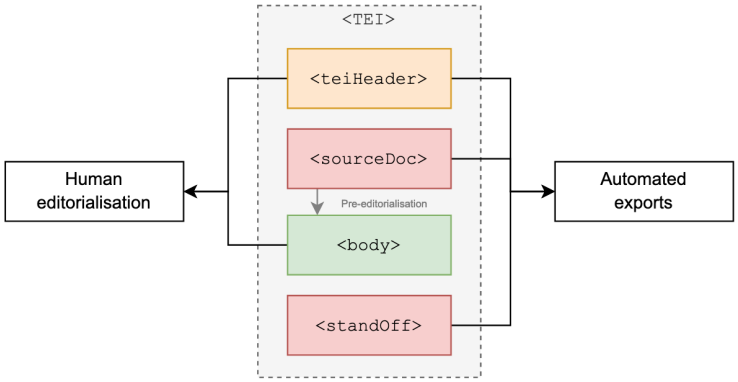

Figure 2: *Gallic(orpor)a* TEI model. In red strictly coercive automatically generated data. In green fully customisable data. In orange partially customisable data.

Data and Script

Data and scripts are available online: `https://github.com/Gallicorpora`.

References

Bartz, A. et al. (Oct. 2021a). "Expanding the content model of annotationBlock". In: *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. Virtual, United States. URL: https://hal.archives-ouvertes.fr/hal-03380805.

— (Oct. 2021b). "Expanding the content model of annotationBlock". In: *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. Virtual, United States. URL: https://hal.archives-ouvertes.fr/hal-03380805.

Bawden, R. et al. (June 2022). "Automatic Normalisation of Early Modern French". In: *LREC 2022 - 13th Language Resources and Evaluation Conference*. European Language Resources Association. Marseille, France. DOI: 10.5281/zenodo.5865428. URL: https://hal.inria.fr/hal-03540226.

Bermès, E. (Jan. 25, 2020). "Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)". These de doctorat. Paris, Ecole nationale des chartes. URL: http://www.theses.fr/2020ENCP0001.

Gabay, S., J.-B. Camps, 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.

Gabay, S., T. Clérice, et al. (Oct. 2020). "Standardizing linguistic data: method and tools for annotating (pre-orthographic) French". In: *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20)*. Hammamet, Tunisia. DOI: 10.1145/3423603.3423996. URL: https://hal.archives-ouvertes.fr/hal-03018381.

Gabay, S., P. O. Suarez, et al. (June 2022). "From FreEM to D'AlemBERT". In: URL: https://hal.inria.fr/hal-03596653.

Gray, J. (2009). "Jim Gray on eScience: A transformed scientific method". In: *The fourth paradigm: Data-intensive scientific discovery*. Ed. by T. Hey, S. Tansley, and K. Tolle. Washington, pp. xvii–xxxi.

Janès, J. et al. (Dec. 2021). "Towards automatic TEI encoding via layout analysis". In: *Fantastic future 21, 3rd International Conference on Artificial Intelligence for Librairies, Archives and Museums*. Paris, France: AI for Libraries, Archives, and Museums (ai4lam). URL: https://hal.archives-ouvertes.fr/hal-03527287.

Pinche, A. and T. Clérice (Aug. 2021). *HTR-United/cremma-medieval: 1.0.1 Bicerin (DOI)*. Version 1.0.1. DOI: 10.5281/zenodo.5235186. URL: https://doi.org/10.5281/zenodo.5235186.

Scheithauer, H. et al. (Oct. 2021). "From page to content – which TEI representation for HTR output?" In: *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. Weaton (virtual), United States. URL: https://hal.archives-ouvertes.fr/hal-03380807.

Biographies

1. Ariane Pinche is a postdoctoral fellow at the École nationale des chartes | PSL and currently works on Medieval manuscripts and HTR in *CREMMAlab* and *Gallic(orpor)a* projects.

2. Kelly Christensen is an intern at Inria and a member of the *Gallic(orpor)a* project. She holds a PhD in musicology and is specialised in digital humanities.

3. Simon Gabay is a senior lecturer and researcher at the university of Geneva. He is specialised in Early Modern French literature and participates in the *Gallic(orpor)a* project.