

---

# Crowdsourcing, scholarly editing and TEI

Wioletta A. Miskiewicz\*<sup>1</sup> and Pierre Willaime\*<sup>1</sup>

<sup>1</sup>CNRS/Archives Henri Poincaré Nancy – CNRS : UMR7117 – France

## Abstract

As a digital editing of the sources, digital humanities call the *natural* materiality of its sources (archives, critical editions, collections...) and consequently their *objectivity* into question. In recent years, owing to proliferation of fakes and hoaxes, the crowdsourcing editing of the humanities archives revive the question of the scientific legitimacy of the *digital scholarly editions*.

In what way can we grant to non-scientists the access to scientific databases without threatening the projects ? To what extent can we allow enlightened amateurs and onlookers to participate to their construction ? And what are the reasons to do that ?

Being convinced that in the new digital ecosystem of science, the scientific legitimacy can be justified only by the construction of enhanced databases, we propose to organize a meeting on the question of the potential that the TEI semantical encoding represents in that context.

We are mainly going to refer to two projects (both projects work in progress): Henri Poincaré's letters (<http://e-hp.ahp-numerique.fr>, and the Archives éLV (<http://www.elv-akt.net/index.php?&langue=en>). It's from that last project that we'd like to develop, in an eventual collaboration with another DARIAH project, a new platform of crowdsourcing online transcription. The platform which will be created for the needs of transcription of a philosophical archive (typescripts, manuscripts of research and lectures, letters and diary) faces increasingly the problems tied to multilingualism and therefore with the question of access to the content in the rare languages archives.

We'd like to give some special attention to the project *Testaments de Poilus* ( Testaments of the French troops of WWI ) <https://testaments-de-poilus.huma-num.fr/#!/> directed by E. de Champs (University of Cergy) and F. Clavaud (National Archives) in order to provide experience feedback. This project, whose platform went online in January 2018, allows registered users who have followed a tutorial, the transcription of testaments of the French soldiers died during WWI/The Great War into XML/TEI tags. Emanuelle de Champs will be our guest during this meeting.

Collaborative transcribing is difficult and challenging. On one side, the transcribing interface must be user friendly to allow a massive participation. On the other side, standards of the semantic web such as RDF should be respected to ensure further scientific exploitation.

There is a gap between these two aspects. Projects have made different choices to fill it. The Bentham project (<http://transcribe-bentham.ucl.ac.uk/>) uses MediaWiki and has developed few extensions for TEI markup support. Testament de Poilus (<https://testaments-de-poilus.huma-num.fr/>) uses Ace Editor library as a frontend for XML TEI. The Consortium

---

\*Speaker

”Archives des ethnologues” offers a different approach with <http://transcrire.huma-num.fr/>. They use Omeka Classic with scripto’s plugin, an integration of MediaWiki adapted for transcripts. This solution without TEI takes advantage of Omeka Classic interface and allows users to indicate metadata for each document.

In the same spirit, The Archives Henri Poincaré (Nancy, France) uses Omeka S to enter and display metadata about Poincaré’s letters (<http://e-hp.ahp-numerique.fr/>). Transcripts of the letters are available in html format and hyperlinks are used to enhance the reader experience. The documentation is properly done in RDF inside the metadata while html transcripts use hyperlinks to refer to the metadata.

None of these solutions is perfect. The power of TEI encoding seems to be hard to conciliate with a user-friendly and simple interface needed for crowdsourcing. Omeka Classic or S offers a intuitive interface to write non-TEI transcripts with the help of Scripto. But there is no link between metadata and transcripts (except with manual hyperlinks which is not a very efficient and standard aware solution).

We’d also like to propose to the Dariah community the question of the usage of TEI encoding in crowdsourcing projects linked to arts and humanities archives, in particular the multilingual ones. Based on the French experience we would like, even before the Warsaw meeting, to make a first round of the question within the forum of DARIAH international community.

**Keywords:** crowdsourcing, online transcription, digital scholarly editions, TEI, arts and humanities archives, multilingual archives