EHRI Meeting - July, 4th 2023

Persons attending the meeting:

- Wolfgang Schellenbacher (Vienna Wiesenthal Institute for Holocaust Studies)
- Aneta Plzáková (Masaryk Institute and Archives of the Czech Academy of Sciences)
- Michal Frankl (Masaryk Institute and Archives of the Czech Academy of Sciences)
- Michala Lônčíková (Masaryk Institute and Archives of the Czech Academy of Sciences)
- Mike Bryant (King's College London)
- Maria Dermentzi (King's College London)
- Floriane Chiffoleau (Inria, ALMAnaCH)
- Sarah Beniere (Inria, ALMAnaCH)
- Hugo Scheithauer (Inria, ALMAnaCH)

Topics discussed during the meeting:

- Presentation of the work Sarah did
- Discussion about the annotations (of named entities)
- Discussion about the *pros* and *cons* of using TEI Publisher
- Conclusion of the meeting → creating a *proof-of-concept*

A – Presentation of the work Sarah did

Here is a link to the slides made and presented by Sarah: https://docs.google.com/presentation/d/1jKMVan4A2vLeNHZgCE6hstv6JpKafKssiNX7kfziB0s/edit?usp=sharing

Sarah worked during her internship in the ALMAnaCH team on a new EHRI ODD which she presented to the EHRI team, as well as a set of recommendations. The new ODD will serve as a validation tool to control TEI conformance, have a more refined and regular encoding, to homogenize all EHRI editions, and to enhance FAIRness. ODD specifications are described in the slides.

She also introduced a semi-automated and documented encoding process based on python script, which creates TEI templates (metadata and text) that can be later filled and modified according to documents to be encoded.

B – Discussion about named entity annotation for EHRI

Currently, annotation remains the most arduous and longest task in EHRI editions. ALMAnaCH proposed TEI Publisher as a more sustainable solution for publishing and annotating EHRI documents. EHRI considers TEI Publisher as an experimental idea because they don't have the means or time to immediately switch to TEI Publisher. However, this can be an interesting solution for the annotations, mainly for the tools established by TEI Publisher. Connecting the platform to EHRI vocabularies is an issue, but it would be interesting to make the process easier and more serene, and engineers at ALMAnaCH are working on it.

Already existing EHRI editions could be used to gather a named entity training corpus and train a model specialized on EHRI's domain. The trained model could then be fine-tuned iteratively when more data is annotated. However, a few issues still remain: the training corpus should be larger; EHRI uses the entity "term", which is not recognized in popular pre-trained models available online; multilingualism may or may not be an obstacle to obtain a robust EHRI model. Word variation depending on the language and the context should also be addressed. Such phenomena can be handled by state-of-the-art named entity and disambiguation models (e.g. Terezín and Theresienstadt both refer to the same entity, the city, but the Theresienstadt Ghetto refers to the ghetto located within the city of Theresienstadt between 1941 and 1945. There are not the same entity.)

As it was discussed with the EHRI members during the meeting, Hugo is currently working on the question of annotating with TEI Publisher, connecting to new authority files and using already annotated data for training a NER model; Maria Dermentzi (King's College, London), who was present at the meeting, is also experimenting on NER pipeline for EHRI, which could lead to associated work between the two to overcome any difficulties.

The ALMAnaCH team proposed to use TEI Publisher for EHRI editions as a platform for annotating, extracting named entities, and publication. TEI Publisher annotation GUI allows authorized users to:

- Use a pre-trained NER model (a generic model or a model already fined-tuned on EHRI documents) for automatically annotating named entities in TEI files.
- Correct annotations if necessary.
- Train or fine-tune a NER model on EHRI data. In other words, EHRI could gather over time an ever-growing named entities corpus that can be used to obtain, in the end, a reliable and robust NER model.

This solution would then streamline the annotation process significantly, while making handling big corpora much easier.

C – Discussion about the *pros* and *cons* of using TEI Publisher

The suggestions made by Sarah with her mock-ups have been globally well-received, and the EHRI members seems to not be against the idea of switching to TEI Publisher. They seemed to think of TEI Publisher as an easier solution to maintain in the long term than Omeka and their specifically designed plugin. There are however some legitimate reservations on the technical side. They are worried that some Omeka plugins might be difficult to implement in TEI Publisher, mostly because many various services are used (for the map, for the links to the vocabularies, etc.). It could be interesting, according to them, to have a two-speed approach, by keeping Omeka and having the EHRI editions in TEI Publisher on the side, as a backup solution, especially if maintaining Omeka become too complicated. The main difference is that what we are offering forces them to reconsider the way the content is presented: we are suggesting a horizontal presentation, while they had a vertical one, notably because it is better for reading and understanding content. It would be relevant, if we continue with our TEI Publisher proposition, to implement this idea, to not crowd the user with information. On their side, they are considering rethinking their design, especially by having some user feedback (using polls for instance). On this subject, we also mentioned the fact that while we propose horizontally the text, the index entries, the interactive map and the facsimile, it is also possible to choose to display or not those items, thanks to buttons allowing to suppress or add again the elements if necessary. Mike Bryant was also wondering what could be missing in Omeka and could be found in TEI Publisher, such as full-text search among the editions, which seems to have been an issue with Omeka, and the existence of an API to work more deeply with TEI Publisher.

It was also mentioned that Sarah already planned, as part of her internship, to write specifications for a dedicated TEI Publisher platform, which would contain relevant and useful functionalities for the EHRI editions. Therefore, with this discussion, she can also think about the transformation of essential Omeka plugins for the editions and how to implement them in TEI Publisher.

D – Conclusion of the meeting → creating a *proof-of-concept*

After our discussion and the arguments that were brought forward, the conclusion of the meeting was that, as TEI Publisher could be advantageous in many ways, the solution is to establish a *proof-of-concept* application, that would serve two purposes. First, this creates for them a TEI Publisher prototype, in which we can implement some suggestions and information proposed by Sarah in her specifications. This also give us the occasion to have a more thorough reflection on what can be implemented or not and so, to answer EHRI's questions about whether TEI Publisher would be a better solution and if it would be best to switch completely to it, especially if something concrete has already been created. The *proof-of-concept* will also allow EHRI to examinate the *pros* and *cons* of the front-end, and the XML database used in the backend.

Second, the *proof-of-concept* will also have the annotation interface configured so that EHRI can begin experimenting with this feature for future editions.

The proposed schedule would be feedback by October or November, with a presentation of the *proof-of-concept*. Before that, we can let them know, around September, where we are at and what is our timeline. Michal is thinking of implementing TEI Publisher in the future post-EHRI-3.