CETHEFI corpus (18th century French plays)

- **Corpus or project name** : CETHEFI Project
- **Coverage of the corpus (temporal, spatial)** : 18th century French theater of Paris (excluding the Comédie-Française): theaters of the Fair and Comédie-Italienne.
- **Information on those responsible for maintaining the corpus and participating institutions** : Scientific manager: Françoise Rubellin, professor of literature, Nantes University, director of CETHEFI (*Centre d'Études des Théâtres de la Foire et de la Comédie-Italienne* / Study Center for Fair Theaters and Italian Comedy) - Technical manager: Olivier Aubert, lecturer in computer science, Nantes University
- **URL and/or corpus repository, if available** : a large part is accessible on http://theaville.org/ some others on http://cethefi.org/ciresfi/doku.php
- **Encoding format (if it is not already TEI/XML):** varied (LaTeX, TEI with different conventions, Word) - work is underway to consolidate the entire corpus into a single TEI corpus compatible with DraCor.
- 300 to 500 words on **progress status of the corpus project concerned** and the expected benefits of contributing to a conference

CETHEFI (*Centre d'Études des Théâtres de la Foire et de la Comédie-Italienne* / Study Center for Fair Theaters and Italian Comedy) has been working for 25 years to transcribe the manuscripts of unpublished plays from the Foire and Comédie-Italienne theaters (see "édition et augmentation du répertoire" pp. 21-24 in "*Spectacles et artistes forains XVIIe-XIXe siècles*" DOI 10.34929/k0nw-wq67)

These pieces were the subject of critical editions during several individual or collective projects (theses, student dissertations, research projects). A part (around 250 opera parodies) was transcribed into LaTeX and put online on the site www.theaville.org. 60 other pieces were transcribed into TEI and put online on the CETHEFI website. Around a hundred more are still in word or odt formats, and scattered. Our goal is to build a virtual library of the Fair and Comédie-Italienne theaters, by primarily publishing texts that are still manuscripts.

Paul Fièvre's Théâtre Classique site, from which the majority of the French DraCor corpus comes, contains some plays from the theaters of the Fair and Italian Comedy, which were printed in the 18th century. But it does not include - with one exception *L'Ile du Gougou* contributed by our project - any edition of non-printed handwritten pieces (we have transcribed more than 300).

One of the current challenges is to FAIRize the collected data. In particular, we will consolidate all the texts of the project as TEI documents, defining and using a TEI schema compatible with the DraCor schema.

A subset of the pieces is available in Word format - which notably brings together the work of students who have transcribed pieces that are sometimes difficult to find and read. We have set up a conversion workflow using the Odette application http://fictif.org/odette/ to convert LibreOffice to TEI. This will provide an initial basis, which will then need to be refined/corrected.

We also have a database (http://www.theaville.org/) referencing the tunes used within the texts, with more than 250 plays, encoded in LaTeX with a dedicated style sheet to be able to express information about the used tunes, their origin, the meter of the verses, etc. Indeed, one of the characteristics of the plays in the corpus lies in the importance of the musical tunes mentioned in the plays: the constraints imposed by the monopoly of the Comédie-Française prohibited the presentation of plays in spoken prose. The troupes therefore circumvented this ban by presenting widely sung pieces: this was the beginning of comic opera. The text is sung to well-known tunes called vaudevilles (known tunes to which we put new lyrics).

We defined a TEI diagram based mainly on the Dracor diagram, adding information specific to the project (airs, metrics, etc.). We are currently developing a workflow to automatically convert the majority of LaTeX files to TEI.

We would like to discuss several topics during the workshop, including

- what are the best practices and tools for converting an existing but heterogeneous corpus of pieces, collected over several decades
- how to express in TEI the information related to the music tunes (vaudevilles) which are the heart of the Theaville project (250 pieces) - there are issues of representation of the information, schema definition, interaction with other existing standards such as MEI, etc
- what further works/connections could be envisioned through this contribution to the DraCor project