Pandore: a toolbox for digital humanities text-based workflows

Alrahabi, Motasem
motasem.alrahabi@gmail.com
Sorbonne Université, France

Fedchenko, Valentina
valentina.fedchenko@sorbonne-universite.fr
Sorbonne Université, France

Petkovic, Ljudmila
ljudmila.petkovic@sorbonne-universite.fr
Sorbonne Université, France

Roe, Glenn
glenn.roe@sorbonne-universite.fr
Sorbonne Université, France

Abstract: We present Pandore, an online toolbox for pre- and post-processing of textual data. This freely accessible platform allows humanities students and researchers, with little or no prior knowledge of computer programming, to perform a series of important and recurrent tasks crucial for the editing, enrichment and mining of digital texts.

Keywords: Corpus, XML-TEI, OCR/HTR, Named Entities Recognition, Semantic annotation, Visualization, Graphical user interface.

Over the past few years, the need for a general-purpose text-processing pipeline for digital humanities (DH) research has become increasingly evident. Students and scholars who want to incorporate DH methods into their research practices are often faced with significant obstacles in the acquisition, enrichment and analysis of digital texts and corpora. To address these challenges, we launched a development project in early 2022 aimed at creating an online toolbox for pre- and post-processing of textual data (Cordova et al., 2022). This freely accessible platform, which we now call Pandore, allows humanities students and researchers, with little or no prior knowledge of computer programming, to perform a series of important and recurrent tasks crucial for the editing, enrichment and mining of digital texts.

Recent studies of corpus-based research practices can provide some insight on methods and challenges for open-ended text analysis tools and their user-centred evaluation (Lyding, 2022; Heuwing et al., 2016). In this context, some existing tools[1] allow entry-level users to analyse their own data, such as LancsBox the Lancaster University corpus toolbox[2] (Brezina et al., 2020); the Cortext Manager[3] (Breucker et al., 2016); the web-based CATMA[4] platform (Horstmann, 2020), etc. Our project differs from these existing projects in several ways: first, Pandore proposes a unified and modular online environment covering the essential pre- and post-processing functionalities needed to capture, structure and enrich digital corpora for later analysis; second, as an open-source project, new functionalities can be easily incorporated either locally or via the DH community; and finally, the tool will be maintained and updated as a core component of our teaching and research activities (Alrahabi et al., 2022). In addition, the low barrier of entry for end users with little or no programming experience, makes it accessible to a large range of researchers, teachers and students across the arts and humanities disciplinary landscape.

Some of Pandore's tools have been developed in-house by the team, while others incorporate open-source and freely available codebases. In both cases, we have created new Python scripts to facilitate the use of these tools through an online graphical interface[5]. Users can upload texts and choose the text-processing workflow most relevant to their research or teaching questions. Currently, Pandore includes the follow tools:

- Extract texts from the Wikisource website[6]: the user can specify the URLs of texts to be harvested, or create a random corpus. It is possible to adjust the size of the corpus to be collected or a percentage of the text to be extracted from a given source.
- OCR / HTR of documents via Kraken[7] or Tesseract[8]. The Tesseract tool is available via the Pandore interface, while Kraken can currently only be accessed *via* the Google Colab scripts[9] documented in our GitHub repository (with the possibility of running them either through the Pandore interface or the local instance of an interface eScriptorium[10] in the future).
- Post-OCR / HTR correction (JamSpell[11]): multilingual contextual spell checking and correcting library, based on 3-gram language models.
- Conversion of files (.txt or .odt) to a TEI-XML format.
- Named-entity recognition in TEI-XML files, with SpaCy[12] or Flair[13].
- Text annotation via Textolab rule-based tool[14].
- Geolocation and mapping of place names in texts (Tanagra[15]), visualisation of character networks in novels (Minerva[16]).

We aim to incorporate new features and methods, such as the adaptation of an existing sequence alignment tool; the standardisation of French orthography in historical texts from the 16th to the 19th centuries; and the generation of vectorized word-embeddings using current NLP work on large language models. We also aim to create detailed documentation in the form of video capsules and tutorials that can be used in teaching.

In addition to the current web interface, we would like to offer the capacity for advanced users to execute and save workflows offline. This is particularly relevant given the current physical limitations for processing large amounts of data directly online, by transforming its current codebase into a suite of libraries executable offline.

Pandore aims to facilitate new practices in, and practice-led reflection on, the application of digital methods to existing humanistic textual sources, demonstrating both the effectiveness of such methods and, at the same time, their limitations. This type of development therefore contributes to the current interdisciplinary dialogue between researchers in Natural Language Processing and the Digital Humanities, opening up a space for knowledge exchange on the role of digital technologies in text-based arts and humanities disciplines.

Notes

1. See, for example, the various DH tool portals available online: https://tapor.ca/home; http://multital.inalco.fr; https://www.ortolang.fr/market/tools, etc.
2. http://corpora.lancs.ac.uk/lancsbox/index.php
3. *https://docs.cortext.net/*
4. *https://catma.de/*
5. *https://obtic.sorbonne-universite.fr/developpement/toolbox/*
6. https://wikisource.org/
7. https://kraken.re/master/index.html
8. https://github.com/tesseract-ocr/tesseract
9. https://colab.research.google.com/github/obtic-scai/Toolbox/blob/dev/OCR/kraken/Kraken_LP.ipynb
10. https://gitlab.com/scripta/escriptorium/
11. https://github.com/bakwc/JamSpell
12. https://spacy.io/
13. https://github.com/flairNLP/flair
14. https://github.com/obtic-scai/Textolab/tree/dev
15. https://obtic.sorbonne-universite.fr/tanagra/map
16. *https://github.com/obtic-scai/Toolbox/tree/dev/Fouille*

Bibliography

Alrahabi, M. / **Roe, G.** / **Koskas, Bordry, C.** / **Bordry, M.** / **J. Gawley** (2022): Enjeux de l'enseignement des Humanités numériques pour les étudiants en littérature: une expérience de classe, in: *Revue Humanités numériques*, 5. https://journals.openedition.org/revuehn/2775 .

Cordova, J. M. / **Dupont, Y.** / **Petkovic, L.** / **Gawley, J.** / **Alrahabi, M.** / **Roe, G.** (2022): Toolbox: une chaîne de traitement de corpus pour les humanités numériques (Toolbox: a corpus processing pipeline for digital humanities). *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 3: Démonstrations* , pp. 12-14. https://hal.archives-ouvertes.fr/hal-03701464 .

Heuwing, B. / **Mandl, T.** / **Womser-Hacker, C.** (2016): Methods for user-centered design and evaluation of text analysis tools in a digital history project. In: *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology. American Society for Information Science, pp. 1-10.*

Horstmann, J. (2020): Undogmatic Literary Annotation with CATMA. In: Nantke, J., Schlupkothen, F. (eds.). *Annotations in Scholarly Editions and Research: Functions, Differentiation, Systematization* , Berlin, Boston: De Gruyter, pp. 157-176. https://doi.org/10.1515/9783110689112-008 .

Lyding, V. (2022): Open demands for corpus analysis tools - a user-centered study. Dissertation. Friedrich-Alexander-Universität Erlangen-Nürnberg. https://opus4.kobv.de/opus4-fau/frontdoor/index/index/docId/18590 .