Averell a management tool to transform XML/TEI poetic corpora in JSON POSTDATA ontology compliant.

Aitor Díaz Medina[a], Javier de la Rosa[a], Alvaro Pérez[a], Laura Hernández Lorenzo [a] Elena González-Blanco[b], Salvador Ros [a]

[a] *POSTADATA Project. SSCC. Escuela Técnica Superior de Informática, UNED, Madrid, Spain*

[b] *Coverwallet. Madrid, Spain*

Introduction

Digital Humanities have led to new, quantitative and more objective as well as scientific approaches to studying the literary record [1]. The case of Poetry is a special one, as there are much less quantitative studies and tools available for this particular genre. In this sense, the process of obtaining texts and information from available poetic corpora from the Web is a hard and time-consuming one.

To bridge this gap, we have developed a tool called Averell within the frame of the POSTDATA ERC project. This is a Python library and command-line tool to download corpora in different TEI formats and transform them into a single JSON representation (see Figure 1). The generated JSON file is created based on the properties and identities of the POSTDATA-core [2], and POSTDATA-structural [3] ontologies.

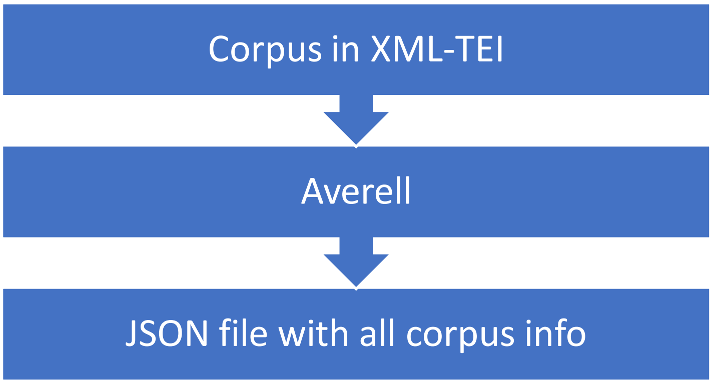

Figure 1. Averell flowchart

This tool has been designed to be useful as a single tool for the DH researcher. Averell is an easy to use tool both from the command line and as python API. Averell will download the corpora from the sources, then parse them, and turn them into a single JSON. This is especially useful for setting benchmarks and to reduce the burden of cleaning and parsing on the researchers. Other functionalities that Averell has are focused on researchers needs. For example, Averell allows researchers compounds new corpora from its catalogue selecting sets of poems from the different corpus in the catalogue. Also, it possible to select the granularity of the new corpora (e.g. stanza, line, syllable). At the moment, its catalogue has five Spanish poetry corpora from different sources.

Finally, we are increasing the number of corpus and languages and studying the extension of its functionalities to other genres as drama, DraCor [4]. We are also working on including more output formats like CSV and RDF formats. The source code is fully available at our Github[5].

Acknowledgements

The authors acknowledge the support of the Starting Grant research project Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528). This project is funded by the European Research Council (ERC) under the research and innovation program Horizon2020 of the European Union.

Bibliography.

[1] Jocker, M. (2013): Macroanalysis. Digital Methods and Literary History. Urbana: University of Illinois Press.

[2]POSTDATA ERC Project. Postadata-Core ontology.

http://postdata.linhd.uned.es/ontology/postdata-core/documentation/index.html

[3] POSTDATA ERC Project. Postdata-structural ontology.

http://postdata.linhd.uned.es/ontology/postdata-structuralElements/documentation/index-en.html

[4] Skorinkin D., Fischer F., Palchikov G. (2018): Building a Corpus for the Quantitative Research of Russian Drama: Composition, Structure, Case Studies. Proceedings of the International Conference "Dialogue 2018", pp. 662–682.

[5] POSTDATA ERC Project: Averell. https://github.com/linhd-postdata/averell