Eurasian Latin Archive

Emmanuela Carbé - University of Siena
emmanuela.carbe2@unisi.it

ELA - Eurasian Latin Archive is a platform under construction aimed at hosting an open access library of Latin and multilingual texts of medieval and early modern age concerning East Asia. The platform will include tools to investigate the documents in their linguistic and semantic aspects.

The start-up phase (March 2018-February 2020) has been cofinanced by Regione Toscana within DAS-MeMo (*Data-mining e analisi statistica su fonti testuali storiche del periodo medievale e moderno*), a project that involves the Department of Philology and Literary Criticism of the University of Siena, along with its Center for Comparative Studies, and the IT Company QuestIT, specialized in Artificial Intelligence and Machine Learning.

The complex and demanding project gives exciting opportunities also from the point of view of the digital humanities studies. It allows to reflect on methodological issues and to seek solutions on a wide range of topics. Challenges starts with the corpus definition, passing through the digitalization/transcription of big amounts of texts, the encoding and the developing of text analysis tools and the automatic extraction of semantic information with Natural Language Processing methods. One of the most interesting test benches of this project concerns the treatment of multilingual texts, on which we are currently working using some excerpts of Intorcetta's *Sapientia Sinica*.

Aim of this paper is to provide an introduction to the project, explaining the analysis of requirements and the general architecture, the reasons of some technical and methodological choices, and the tasks planned in the middle and long terms. The paper will also show the first prototype of the platform (ELA – alpha version), available for this workshop at the URL http://www.dasmemo.it. In the prototype documents are freely searchable bu means of an ElasticSearch search-based search engine developed by dr. Nicola Giannelli (QuestIT). All texts are being encoded in XML TEI, following the guidelines adopted by the ALIM Project (alim.unisi.it), and also include our first experiments on Named-entity recognition.

Keywords
Digital Humanities; Digital Archives; multilingual documents; TEI; NER; NLP