System of Indexing, Annotation and Search in the old Arabic Manuscripts

Noureddine EL MAKHFI[1], Omar EL BANNAY[1], Rachid BENSLIMANE[1], Noureddine RAIS[2]

n.elmakhfi@gmail.com , omarelbannay@gmail.com , r.benslimane1@gmail.com , raissn@gmail.com

[1] *Laboratoire de Transmission et Traitement de l'Information (LTTI), EST*
[2] *Laboratoire d'Informatique Statistique et Qualité (LISQ), FSDM*

Abstract **—** Ancient manuscripts constitute a heritage to preserve for future generations and make it accessible to a wider audience.

This heritage is unfortunately consulted by a minority of experts and researchers because of the lack of information on the existence of this national heritage. The problem is partly due to the fact that this manuscript heritage is stocked in libraries scattered around the country and managed by institutions belonging to different departments or different establishments of the same department. On the other hand, restriction of access to national heritage manuscript is related to the concern to preserve the manuscripts physically manipulated which contribute to their accelerated degradation, taking into consideration these limitations on access while ensuring preservation of original manuscripts, the solution widely adopted by developed countries is based partly on the digitization of this heritage manuscript, and partly on the development of management platforms and diffusion of this wealth of knowledge digitized.

Easy access to such manuscripts presented on images format requires an index which can be created manually or automatically by using Optical Recognition Characters (OCR). The automatic approach is difficult to realize by considering the cursive nature of Arabic writing and by the exaggerated overlapping between words and lines in handwritten Arabic manuscripts. So, the segmentation leading to extract respectively and separately lines, words and characters constitutes the critical operation, affecting the performance of all Arabic (OCR) systems. Recent research works adopt the word spotting approach. The words in a manuscript are matched as images and grouped into clusters which contain all instances of the same word. This new and original approach solve the problem related to the cursive nature of the Arabic writing by not considering the character segmentation, but considers the line and word segmentation perfectly performed.

To tackle these problems mainly connected to segmentation operation, we propose in this paper a new indexing system of Arabic manuscripts, which facilitates transcription and the establishment of Arabic handwritten text by annotating images of manuscripts according to metadata. Our system offers the ability to create works of Arabic manuscripts in a format TEI XML, this model includes all the primitives that are related to the international standard markup TEI (Text Encoding Initiative), and to provide good guarantees compatibility with existing tools. This system supplies the database of the open source platform SDX (System Documentary XML) that allows searching, browsing and viewing of these documents in a web format.

The good performances of the proposed method were tested by using some ancient Arabic manuscripts.

Keywords **— Annotation, Arabic Manuscripts, Digitalization, Handwriting Recognition, Metadata, Transcription**