Digital safeguard of laminated historical manuscripts: the treatise "Poem in Rajaz on medicine" as a case study

Angelo Mario Del Grosso[1], Driss Fassi Fihri[2], Mohammed El Mohajir[3], Ouafae Nahli[1] and Anna Tonazzini[4]
[1]Institute for Computational Linguistic, Italian National Research Council, Pisa, Italy
Email: ouafae.nahli@ilc.cnr.it, angelo.delgrosso@ilc.cnr.it
[2]University of Al-Qarawiyyin, Fez, Morocco
Email: fassifihridriss@hotmail.com
[4]Sidi Mohamed Ben Abdellah University, Fez, Morocco
Email: m.elmohajir@ieee.ma
[3]Institute of Information Science and Technologies, Italian National Research Council, Pisa, Italy
Email: anna.tonazzini@isti.cnr.it

***Abstract*—In this paper, we analyze and discuss the characteristics of a system for the effective digital preservation and fruition of historical manuscripts degraded by the process of lamination. As a case study, we will make reference to the "Poem in Rajaz on medicine", written by Abubacer in the XII century, and conserved in the Al Quaraouiyine Library located in Fez, Morocco.**

The conceived system should have at least four main functionalities: image acquisition (i.e. digitization), image enhancement, text encoding, and linguistic analysis. Based on the evaluation of the manuscript damages, the acquisition set up should be designed in such a way to be able to avoid reflections as much as possible. Suitable digital image processing techniques should also be devised to correct the residual degradations and enhance the text for an easier legibility. Finally, semi-automatic transcription, scholarly encoding and linguistic analysis, to be performed on the virtually restored pages, should adapt existing tools to the specificity of the primary source writing system and language.

The feasibility study for the realization of such a system is of general utility, in that it can provide guidelines for the digitization, the enhancement and the text encoding of the many laminated manuscripts conserved in other historical archives. On the other hand, from the cultural heritage point of view, the experimentation on the "Poem in Rajaz on medicine" could foster the systematic philological and ontological study of a unique piece of our documental heritage: the longest poem of medieval Islamic medical literature.

***Index Terms*—Cultural Heritage Digital Safeguard; Historical Manuscript Digitization; Document Image Processing; Linguistic Analysis; Ontological Analysis**

I. Introduction

In recent years, extensive campaigns of digitization of the documental heritage preserved in libraries and archives have been performed, with the primary goal to ensure the safeguard and the fruition of this important part of the human cultural and historical legacy. Besides ensuring conservation against future damages, the availability of high quality digital surrogates has increasingly stimulated the use of image processing techniques, to perform a number of operations on documents and manuscripts, without harming the often precious and fragile original sources. Among those, virtual restoration tasks are crucial for attenuating degradations suffered during time, and improving legibility of the text of interest. Automatic or semi-automatic processing of the digital images can also be performed with the purpose of extracting the information necessary to some downstream tasks, such as textual analysis, transcription and annotation. Finally, software tools for linguistic analysis exist that build advanced representations of the information content of the manuscripts through text processing at different levels of complexity: morphological analysis, syntactical analysis, and semantic interpretation. All the instruments mentioned above help paleographers and philologists in their work, thus facilitating a more exhaustive, complete and efficient preservation of the legacy that historical manuscripts hold. In other words, digital safeguard of historical manuscripts can be considered in a wider sense, which overpasses the acquisition and the proper storage of the digital images alone, and includes also linguistic analysis and comprehension of the written contents.

In this paper, we discuss the possible architecture and feasibility of a complete system for the digital safeguard of historical manuscripts degraded by the process of lamination. As a case study, we will make reference to a very important Moroccan manuscript, the "Poem in Rajaz on medicine" (from now on, the Poem), written by the physician and philosopher Abubacer in the XII century [1] [2]. This is the longest poem of medieval Islamic medical literature. Considered as an encyclopedia of diseases and treatments, but also of botany and zoology, it is especially important in the study of folk and herbal medicine, history of the medical evolution between pharmacy and chemistry, history of diseases and drugs, and recognition of the disease symptoms [3]. Unfortunately, only one copy of this manuscript exists, and it is conserved within the Al Quaraouiyine Library located in Fez, Morocco, where it is catalogued under the number 3158/0040 [4]. Unluckily,

in the 1960's, the manuscript has been laminated, and this process caused it severe damages. The poor conservation state of this manuscript is similar to that of many other historical manuscripts in the libraries, archives and museums of all countries. The concrete risk that the irreversible operation of lamination will soon make those manuscripts unreadable imposes the urgency of their digital preservation, analysis and philological study. To this purpose, an effective computational system could be inspired to the plan that Madani Salih proposed for the digital safeguard of the Poem itself [5]:

i) digitization of the manuscript; ii) digital image processing to have a clearer version, facilitating reading; iii) digital transcription of the text; iv) critical edition of the corrected version with comments that explain the text; v) transmission to specialists in folk and herbalist medicine, and specialists in Arabic medicine, who can study and comment the text; vi) translation of the text into English.

According to this schedule, we focus our attention on points i)-iii), which are devoted to the digitization and image elaboration of laminated manuscripts, the transcription and the linguistic analysis of the text. The paper is organized as follows. Section II is devoted to the description and the discussion of effective strategies for the digitization of laminated manuscripts, and the subsequent image processing techniques to be applied for recovering a clearer reading of the text. In Section III we describe our approach to the transcription, text encoding and linguistic analysis, with special attention to the Arabic texts. Section IV concludes the paper.

II. Digitization of laminated manuscripts and enhancement of the images

In the last century it was a common practice to cover ancient or precious manuscripts and drawings with chemical substances producing a sort of semi-transparent plastic coat, in an attempt to stop the degradation process of the materials. Nowadays, it is recognized that lamination is by no means effective for delaying the physical decay, and causes itself serious and irreversible damages to the manuscripts, such as, for instance, the warping of the medium (paper or parchment), and/or changes in the color of the inks.

In addition, the digitization process of a manuscript that has been laminated or covered by a transparent varnish, as that of a painting protected by glass, is particularly challenging due to the phenomenon of light reflection.

Indeed, when we take a picture through a semi-transparent medium, we observe an image that is often a superposition of the image of the object beyond the medium and the image of the scene or of the light source located in front of the object and reflected by the medium. We call transmitted image the ideal image of the object of interest, and reflected image, or reflection, the image of this second object.

Professional photographers use polarizing lenses to reduce the intensity of the reflection. Polarimetric imaging systems [6] [7], which incorporate a polarizer directly in the optics, such as cameras equipped with a liquid crystal polarizer [8], can even totally eliminate reflection [9] [10]. However, this can be achieved under a condition that is difficult to be satisfied, namely that the viewing angle is equal to the Brewster angle [11]. In case of plastic-coated manuscripts, which are locally warped by effect of the lamination, another possible expedient to eliminate or at least reduce reflection is to change location and direction of the light source, depending on the local warping. This, however, implies that a manuscript page must be subdivided in more areas to be acquired separately. Image registration and stitching algorithms must then be employed to recompose the entire page.

Whatever the acquisition set up chosen, images totally free from reflections will be rarely obtained. Subsequent elaborations of the available images are then necessary, to perform a virtual restoration of the manuscripts. The majority of the proposed computational approaches assume that the observed image can be considered as a linear combination of the reflected and transmitted images. That is, the observed image is an unknown linear mixing of two unknown images. This model was derived in [12] by analyzing optical models. Mathematically, the problem of recovering the transmitted image from the observed image is highly ill-posed since also the coefficients of the linear combination are unknown and the number of unknowns is twice the number of equations. A first approach to handle this kind of underdetermined problems is to use blind statistical methods of independent components analysis (ICA). We tested ICA on some of the RGB images of the educational digital version of the Poem [1]. Figures 1(a) and 1(b) show one of the original images, in color and in grayscale, respectively. It is apparent that the strong reflection affecting this non-specialized acquisition clearly masks a good deal of the written content. Figures 1(c) and 1(d) show two components extracted through ICA. These illustrate the effect of separation of text and reflection that we expect by ICA under a condition of diversity of acquisitions. However, in this case of acquisitions under a fixed illumination source, separation cannot be full and, above all, the text masked by the reflected light cannot be recovered.

Better removal of the reflection and text recovering have been obtained by ICA from pairs of images of a same scene acquired with two different polarizations of the light source, as described in [12]. For the same kind of acquisition modality, sparse ICA (SPICA) has also been used [13]. In [14], the physical properties of polarization of a double-surfaced glass medium are exploited within a multiscale scheme, to separate the reflection from the transmitted background scene using three polarized images, each captured from the same viewpoint but with different polarization angles, separated by 45 degrees.

When only a single image is available stricter constraints on the problem formulation or on the component images must be exploited. In [15], for example, the problem is handled using local features, and, in [16], by using priors describing sparsity and user-provided information. The dependency of the color channels of the transmitted image, and their independence from the achromatic reflection, is proposed instead

[1] the pdf is available at https://archive.org/details/ibntofayl-urjuzatibbiy

(a) (b)

(c) (d)

Fig. 1. Elaboration through ICA of a page of the Poem: (a) the original RGB image; (b) the original grayscale image: (c) one ICA component showing the map of the reflected light; (d) another ICA component showing the text free of reflections. These results are for illustration purposes only, the poor quality is due to the low resolution of the acquisition and the pdf compression of the recreational images used.

as constraint in [17], where a Maximum A Posteriori (MAP) estimation approach is adopted, which takes also into account the regularity of the images and the differences in their structures [18]. In [19], assuming the same modelling as in [17], we enforced the coincidence of the gradients of the three color channels of the normalized transmitted image, and the statistical independence of those gradients from the gradient of the normalized reflected image. We then proposed a very fast algorithm constituted of two subsequent steps, both based on the above constraints. The first step estimates the model parameters through an ICA algorithm. The second step, based on the now determined data model, estimates the four component images via regularization techniques.

With this approach, when the reflection only partially masks a part of the written content, we obtained results of the kind of those shown in Figure 2. In particular, Figure 2(a) shows the original RGB image, and Figures 2(b) and 2(c) show the two components extracted [19]. For the Poem (in the digital educational images), the reflected light completely masks the text, since no specialized acquisition setup has been exploited. Hence, the same method was not able to recover the masked text, in that case.

Based on the survey above, it is likely that an operative protocol, possibly simple and low-cost, to effectively digitize laminated manuscripts could consists in performing multiple acquisitions based on different light polarizations, and then applying ICA-like algorithms for obtaining the full separation of transmitted and reflected images. We plan to test such a strategy on selected pages of the Poem. In a natural way, validation and performance evaluation of digitization and image enhancement will be based on the comparison between transcriptions carried out from images obtained with standard imaging and with specialized imaging.

III. Text encoding and linguistic analysis

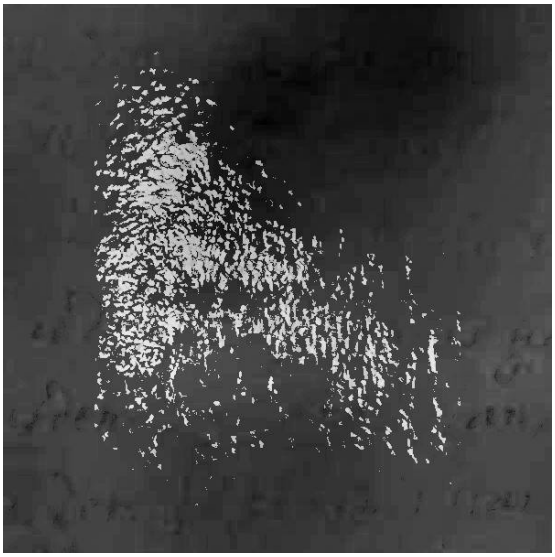
A. Web-based System

The conceived system provides scholars with text-based functionality such as text encoding management and linguistic analysis of the content conveyed by the virtually restored pages. Indeed, one of the purposes of image processing is typical to improve legibility - understood as the ease of the reading, transcription, and linguistic analysis of the text contained in the manuscript.

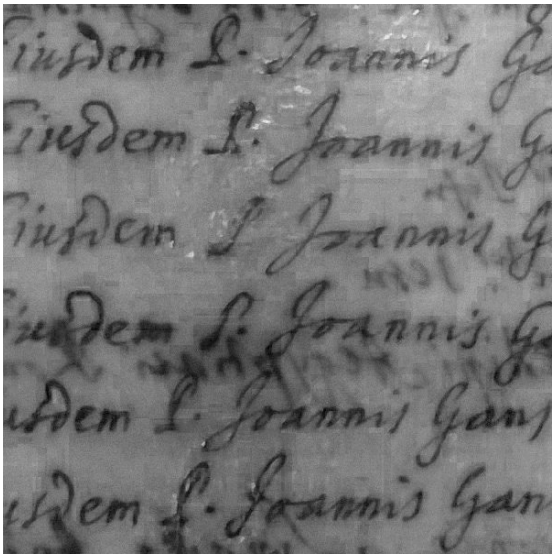
In order to scholarly study and edit primary sources, a computational system has to provide effective support for encoding and for processing the digital representation of textual content. To accomplish this latter task, our system combines tools for analyzing Arabic texts taking into account morphological, syntactical, semantic and philological perspectives.

Concerning the software architecture, we have at our disposal a home-designed platform made up of decoupled and extensible but interrelated components that, following the micro-services approach, are able to accommodate new features and/or different types of text processing, with particular attention to Arabic language.

(a)



(b)



(c)

Fig. 2. Elaboration through the method in [22] of a manuscript image affected by reflection: (a) the original RGB image; (b) the map of the reflected light; (c) the text free of reflections.

Thereby, the edition of the historical manuscripts is based on the use of state-of-the-art methods in digital scholarly editing [24], and on the progressive publication by means of an interactive open source web platform. This under-construction platform in going to provide: a) advanced management functions (zoom, rotation, 3d, YUV) for high-resolution facsimile reproductions (300 to 600 DPI); b) indexing and consultation functions about gazettes and lists of named entities (person names, place names, analytical and iconographic indexes); c) advanced search functions on orthographic and linguistic basis [20] [21]. In addition, the system has to handle: 1) a collection of IIIF-compliant images, 2) the transcription of textual content via an online editor, 3) the text investigation by means of advanced search and indexing features, 4) the encoding of the annotation about the primary source with a special regard to the structure and the content of the original document.

B. Scholarly Text Encoding and Linguistic Analysis

Text - especially literary and/or historical one - is much more than a mere sequence of ordered characters. Indeed, such resources are complex objects with semantic structures conveying multiple meanings and subject to multiple interpretations. In light of all this, digital surrogates of textual documents have to take into account such a multidimensional nature of their primary sources (e.g. physical, logical, historical, linguistic, communicative etc) and therefore they have to make explicit, machine readable and machine actionable all these aspects. A formal and shared model to scholarly encode texts and to digitally record linguistic analyses is the model defined by the Text Encoding Initiative (TEI) [25]. [2] Currently, the TEI model is implemented as an XML schema providing a wide vocabulary with more than 550 elements and more than 250 attributes meeting almost all scholarly needs. The TEI project is well organized in a modular framework, allowing effective customization and extension to adapt it to every scholar's requirements.

Furthermore, linguistic analysis is an advanced representation of the information content of the documents through text processing at different levels of complexity: morphological analysis, syntactical analysis, and semantic interpretation. With specific reference to our case study, since the written Arabic does not contain vowels, lemmatization requires, first of all, the vocalization of words. This step is very demanding and entails a responsibility, as it includes "interpretation" and understanding of the text [22]. In addition, the Poem represents a corpus of medical domain, from which a new interesting terminological network could be extracted. Terms of medical domain have to be extracted manually by the expert, who is in charge of identifying: i) the relevant concepts, e.g. the anatomic structures, but also the concepts related to the diagnosis, the prognosis, etc. ii) the semantic relationships between them.

[2]The TEI encoding scheme is currently the de-facto standard to encode historical manuscripts within a philological perspective.

Figure 3 shows an XML-TEI fragment as a representative example of some relevant features we considered for scholarly encoding the *'urjūzah*. The encoding schema [3] represents the TEI conforming document for our case study showing the two principal blocks 1) metadata and 2) text structure. The main hierarchical relationships we have encoded are as follows:

- `<div1>` which represents the original Sections i.e. the *maqālah* in the *'urjūzah*. They are seven sections, each of them discusses a part of the body, the symptoms of the diseases that can be observed and finally the treatments. Figure 3 presents the encoding of Section 03 which begins on page 48 recto.
- `<div2>` which represents the original Chapters (*bāb*). Each section is divided into several chapters. For example, Figure 3 illustrates the Chapter which begins on page 62 verso and discusses *the indigestion and lack of appetite* in fifteen verses. We noticed that the therapies are announced via a sub-title that has been marked through `<fw>` element.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TEI SYSTEM "tei_all.dtd">
<TEI>
    <teiHeader>…
    </teiHeader>
    <text>
        <body>
            <div1 xml:id="mq03" type="section">
                <pb xml:id="p48r" n="48r" facs ="#facs-p48r"/>
                <head>
                    <title>المَقَالَةُ الثَّالِثَةُ</title>
                    <abbr>3</abbr>
                </head>
                <!-- other div2s -->
                <div2 xml:id="bab017" type="chapter" n="17">
                    <pb xml:id="p62v" n="62v" facs ="#facs-p62v"/>
                    <head>
                        <title>بَابُ سُوءِ الهَضْمِ وَبُطْلَانِ الشَّهْوَةِ</title>
                        <abbr>17</abbr>
                    </head>
                    <!-- other lines -->
                    <l xml:id="bab017-15" n="bab 17 line 15">…
                    </l>
                    <p xml:id="bab017-t1" facs="#facs-t1">
                        <fw type="therapy">عِلَاجُ سُوءِ الهَضْمِ</fw>
                    </p>
                    <l xml:id="bab017-t1-1" n="bab 17 therapy 1 line 1"> …
                    </l>
                    <!-- other lines -->
                </div2>
                <!-- other div2s -->
            </div1>
            <div1 xml:id="mq04" type="section">…
            </div1>
        </body>
    </text>
</TEI>
```

Fig. 3. Example of encoding the hierarchical structure of the poem. the div1 elements represent the poem sections (), the div2 elements represent the poem chapters (), the l elements represent the poem verses.

Figure 4 shows various kinds of TEI elements to encode 1) verse lines by using the `<l>` element; 2) the general purpose `<seg>` element has been adopted to encode the

[3]The XML-TEI schema presented in figure 3 is a customization of the encoding model adopted within the Musisque Deoque project (MQDQ project, http://mizar.unive.it/mqdq/public/), which the authors are involved in. The project, led scientifically by prof. Paolo Mastandrea and technically by Luigi Tessarolo, is one of the broadest and most authoritative digital archives of Latin poetry. In order to encode our case study we used different TEI modules, among which Module 6 (verse), Module 10 (manuscript description), Module 11 (primary sources), Module 16 (segmentation) and Module 17 (linguistic analysis)

two hemistiches for each verse (right and left, respectively); 3) orthographic tokens, which are contained within verses by using the <w> element.

Some relevant tokens which can be referred to the lexicon of the domain are enriched with linguistic and terminological annotations. Particular attention has been reserved for polyrematic terms of the medical domain. For example, the term *"indigestion"* is expressed in Arabic through two words. The initial word *sū'* which corresponds to the prefixes *'in-'* or *'mis-'* and the second word *al=haḍm* which means *'digestion'*. Both of these two words have been linked to the term *"indigestion"* and annotated with the attribute `@part` which means that the selected term starts with the token having such an attribute value "I" (Initial) and ends with the token having this attribute value "F" (Final).

```
<p xml:id="bab017-t1" facs="#facs-t1">
    <fw type="therapy">عِلَاجُ سُوءِ الهَضْمِ</fw>
</p>
<l xml:id="bab017-t1-1" n="bab 17 therapy 1 line 1">

    <seg type="hemistich-right" xml:id="h-bab017-t1-1-left">
        <w xml:id="token07">فَا نْظُرْ</w>
        <w xml:id="token08">فَإِنْ</w>
        <w xml:id="token09">رَأَيْتَ</w>
        <w xml:id="token10" type="domain"
            lemma="سُوء" pos="Noun" lemmaRef="#indigestion" part="I">سُوءَ</w>
        <w xml:id="token11" type="domain"
            lemma="هَضْم" pos="Noun" lemmaRef="#indigestion" part="F">هَضْمِ</w>
    </seg>

    <seg type="hemistich-left" xml:id="h-bab017-t1-1-right">
        <w xml:id="token12">تَبْتَغِي</w>
        <w xml:id="token13" type="domain"
            lemma="عِلَاج" pos="Noun" lemmaRef="#therapy">عِلَاجَ</w>
        <w xml:id="token14">هَذَا</w>
        <w xml:id="token15" type="domain"
            lemma="سَقْم" pos="Noun" lemmaRef="#sickness">السَّقْمِ</w>
    </seg>
</l>
<!-- other lines -->
```

Fig. 4. Example of encoding the therapy block of verses and the linguistic annotations along with single and polyrematic terms.

IV. Conclusions

We have depicted the architecture and the functionalities of a web-based system devoted to the digital preservation and fruition of the large amount of documental heritage that is in degraded conditions due to the lamination process often carried out on the manuscripts in the last decades. We analyzed feasible digitization strategies able to reduce, as much as possible, the light reflection phenomenon, due to both the reflectivity characteristics of the plastic coat and the warping of the support. Subsequent image processing techniques have been revised that can eliminate the residual reflection from the acquired images. The new acquisitions and subsequent digital elaborations allow a better readability of the text conveyed by the document, thereby supporting the process of transcription, text encoding and linguistic analysis through specialized automatic or semi-automatic computational tools.

We have evaluated the feasibility of the proposed method on same pages of "Poem in Rajaz on medicine" manuscript preserved within the library of the Al Quaraouiyine Library located in Fez, Morocco.

References

[1] "The World of Ibn Tufayl: Interdisciplinary Perspectives on Hayy ibn Yaqzan", Lawrence I Conrad (Ed.), Islamic Philosophy, Theology and Science, Texts and Studies, vol. 24, Leiden and New York, E J Brill, 1996.

[2] In Arabic poetry, the Rajaz Meter - the simplest and most common - has been widely used to create mnemonic works to facilitate the memorization of key points and arguments on a given topic. In fact, educational nature and style clarity of the Ibn ufayl's urguzah shows his educational side.

[3] 1980, pp.75-80

[4] al-Qarawiyyin, manuscript location

[5] Madani Salih

[6] T. Cronin, N. Shashar, and L. Wolff, "Portable imaging polarimeters", in Proc. ICPR 1994, Vol. A, pp. 606–609.

[7] K. Nayar, X. Fang, and T. Boult, "Separation of reflection components using color and polarization", Int. J. Comput. Vis. 21, 163–186 (1997).

[8] H. Fujikake, K. Takizawa, T. Aida, H. Kikuchi, T. Fujii, and M. Kawakita, "Electrically-controllable liquid crystal polarizing filter for eliminating reected light", Opt. Rev. 5, 93–98 (1998).

[9] Y. Schechner, J. Shamir, and N. Kiryati, "Polarization based decorrelation of transparent layers: the inclination angle of an invisible surface", in Proc. ICCV 1999, pp. 814–819.

[10] Y. Schechner, J. Shamir, and N. Kiryati "Polarization and statistical analysis of scenes containing a semireflector", J. Opt. Soc. Am. A 17, 276–284 (2000).

[11] M. Born and E. Wolf, Principles of Optics (Pergamon, London, 1965).

[12] H. Farid and E. Adelson, "Separating reflections and lighting using independent components analysis", in Proc. CVPR 1999, Vol. 1, pp. 262–267.

[13] A. Bronstein, M. Bronstein, M. Zibulevsky, and Y. Zeevi, "Sparse ICA for blind separation of transmitted and reflected images," Int. J. Imag. Syst. Technol. 15, 84–91 (2005).

[14] N. Kong, Y.-W. Tai, and J. Shin, "A physically-based approach to reflection-separation: from physical modeling to constrained optimization", IEEE Trans. Pattern Anal. Mach. Intell. 36, 209–221 (2014).

[15] A. Levin, A. Zomet, and Y. Weiss, "Separating reflections from a single image using local features," in Proc. ECCV 2004, pp. 306–313.

[16] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior", IEEE Trans. Pattern Anal. Mach. Intell. 29, 1647–1655 (2007).

[17] K. Kayabol, E. Kuruoglu, and B. Sankur, "Image source separation using color channel dependencies" in Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation, pp. 499–506, 2009.

[18] Q. Yan, E. E. Kuruoglu, X. Yang, Y. Xu, and K. Kayabol, "Separating reflections from a single image using spatial smoothness and structure information", in Proc. LVA/ICA 2010, LNCS, Springer, 2010, Vol. LNCS 6365, pp. 637–644.

[19] L. Bedini, P. Savino, and A. Tonazzini, "Removing achromatic reections from color images with application to artwork imaging", in Proc. 9th IEEE ISPA 2015, pp. 126–130.

[20] A. M. Del Grosso, A. Bellandi, E. Giovannetti, S. Marchi and O. Nahli, "Scanning is Just the Beginning: Exploiting Text and Language Technologies to Enhance the Value of Historical Manuscripts", ISBN 978-1-5386-4385-3, IEEE-CIST 2018 LED-ICT, Marrakech, Morocco, 21-27/10/2018, published by IEEE, New York. Pages 2014-220.

[21] A. M. Del Grosso; O. Nahli, "Towards a flexible open-source software library for multi-layered scholarly textual studies: An Arabic case study dealing with semi-automatic language processing", in: Proc. IEEE CIST 2014.

[22] O. Nahli, "Computational contributions for Arabic language processing Part I. The automatic morphologic analysis of Arabic texts", in Studia graeco-arabica, Pacini Editore, Pisa (Italia).

[23] O. Nahli, S. Marchi, "Improved Written Arabic Word Parsing through Orthographic, Syntactic and Semantic constraints", in Proc. CLiC-it 2015, Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto Eds., Accademia University Press 2015, pp. 210-214.

[24] E. Pierazzo. "Digital Scholarly Editing: Theories, Models and Methods". Farnham, Surrey: Ashgate, 2015. x, 242 p., ill. ISBN 978-1472412119.

[25] L. Burnard, 2014. "What is the Text Encoding Initiative? How to add intelligent markup to digital resources". Marseille: OpenEdition Press 2014.