Challenges in the digital preservation of historical laminated manuscripts

Angelo Mario Del Grosso[1], Driss Fassi Fihri[2], Mohammed El Mohajir[3], Anna Tonazzini[4] and Ouafae Nahli[1]
[1]Institute for Computational Linguistics, 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
[3]Sidi Mohamed Ben Abdellah University, Fez, Morocco
Email: m.elmohajir@ieee.ma
[4]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.**

The most significant degradation caused by lamination is that the parchment or paper support loses its flatness, and usually presents ripples and warnings. This, together with the affixed translucent varnish, dramatically impair the digital acquisition process, so that light reflections in the more disparate directions affect the digital images.

A digital system to contrast this irreversible and progressive degradation and to enable an effective access to the fragile asset should provide a number of functionalities: specialized digitization, able to avoid reflections as much as possible; image enhancement, devised to correct the residual degradations and enhance the text for an easier legibility; semi-automatic transcription of the virtually restored pages; and, finally, scholarly encoding and linguistic analysis, which should adapt existing tools to the specificity of the primary source (writing system and language).

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 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 of ensuring 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 devoted to 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 kept in historical manuscripts. 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 الأَرجُوزَة الطِّبِّيَّة "Poem in Rajaz on medicine" [1] (from now on, the Poem), written by the physician

[1]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 Ţufayl's ʾurğūzah shows his educational side [4].

and philosopher Ibn Ţufayl [2] 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. 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*.

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. Since the lamination process is an irreversible operation, the concrete risk is those of a complete loss of the text readability. Consequently, the urgency imposes a digital preservation of these manuscripts, together with a philological study and a textual analysis of the content.

To this purpose, an effective computational system could be inspired by the plan that Madani Salih proposed for the digital safeguard of the Poem itself [4]:

1) digitization of the manuscript;
2) digital image processing to have a clearer version, facilitating reading;
3) digital transcription of the text;
4) critical edition of the corrected version with comments that explain the text;
5) transmission to specialists in folk and herbalist medicine, and specialists in Arabic medicine, who can study and comment the text;
6) translation of the text into English.

According to this schedule, we focus our attention on the above points 1-3, which are devoted to the digitization, image processing, scholarly transcription, and linguistic analysis of the text.

The paper is an extension of our conference paper [25], where a first sketch of the system was proposed. Here, we describe our ideas in more detail, and added the experimentation, on a small part of the Poem, of the various functionalities of the system.

The paper is organized as follows. Section II is devoted to the description 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. We also discuss those strategies in relationship to the specificity of the Poem. In Section III we describe our approach to the transcription, text encoding and linguistic analysis of the Arabic texts. In particular, we processed a small part of the Poem, which required special attention to the digitization step, and concerns "indigestion, lack of appetite and care". Finally, 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 the warping of the medium (paper or parchment), and/or changes in the color of the inks.

In addition, the digitization of a manuscript that has been laminated or covered by a transparent varnish, as well 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 [5] [6], which incorporate a polarizer directly in the optics, such as cameras equipped with a liquid crystal polarizer [7], can even totally eliminate reflection [8] [9]. 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 [10]. 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.

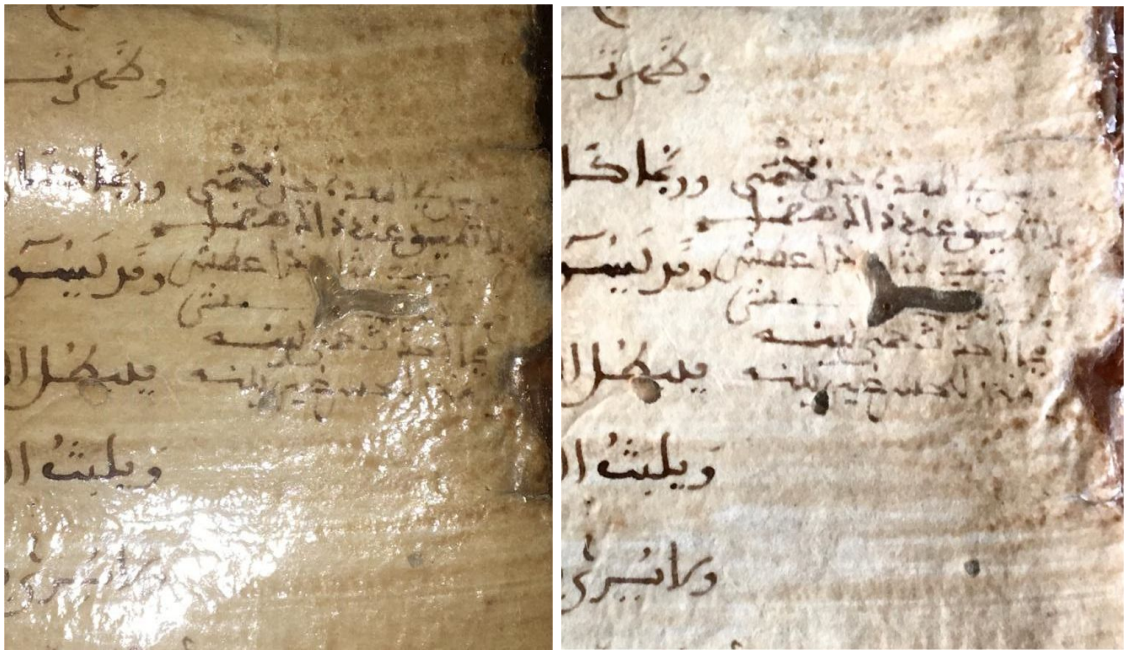
An example of such a strategy, applied to a portion of the Poem, is given in Figure 1. The image in Figure 1(a) has been acquired in a non-professional way,[3] the illumination causes a large reflection, which impairs legibility. Whereas, the image shown in Figure 1(b) is acquired with a proper illumination setup and the camera used in digitizing the document has been a professional device available to the historical archive. In this case, the improvement in readability is impressive.

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. Such a process is very demanding and computational expensive, as requires a significant human intervention, so that it can be adopted for a limited number of samples.

Whatever the acquisition setup chosen, images totally free from reflections will be rarely obtained. Subsequent elaborations of the available images are then necessary, to perform

[2]Two works of Ibn Ţufayl only have come down to us. The first, "risālatu ḥayy ibn yaqān fī al=ḥikmat al=maŝriqiyya", is considered the first philosophical novel in the history of literature. It describes the self-taught education and progress in the knowledge of a human being living alone in an uninhabited island. In fact, Ibn Ţufayl is considered the precursor of Daniel Defoe and Rudyard Kipling. the second work, "Poem in Rajaz on medicine", reflects Ibn Ţufayl's medical career.

[3]The image in Figure 1(a) is captured from the pdf of an educational version of the Poem available at https://archive.org/details/ibntofayl-urjuzatibbyi.



(a) (b)

Fig. 1: Fragment detail of the 62nd page verso of the Poem acquired with two different illumination setups: (a) the illumination causes a large reflection, which impairs legibility; (b) with a properly studied illumination setup no reflection is obtained and the characters are all clearly legible.

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 [11] 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 not known 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), which exploit diversity of acquisitions (e.g. different wavelengths, or different polarizations of the light), and assume statistical independence of the two added images [26].

Satisfactory 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 [11]. For the same kind of acquisition modality, sparse ICA (SPICA) has also been used [12]. In [13], 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.

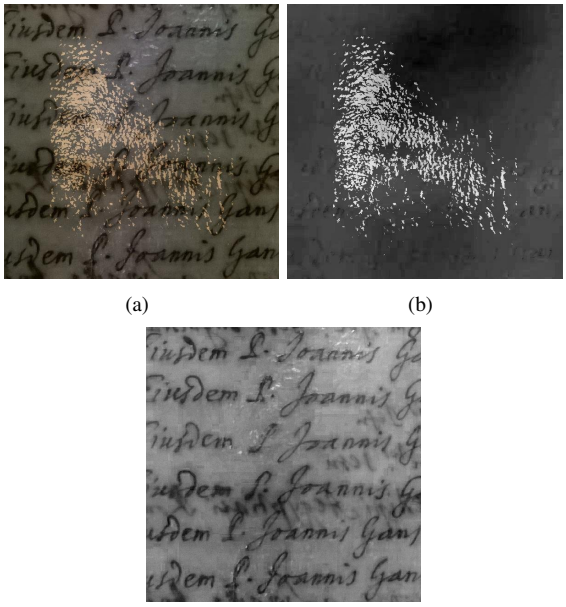
When only a single observation is available, stricter constraints on the problem formulation must be exploited. In [14], for example, the problem is handled using local features, and, in [15], by using priors describing sparsity and user-provided information. For an RGB image, the dependency of the color channels of the transmitted image, and their independence from the achromatic reflection, is proposed [16]. 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 [17].

In [18], assuming the same modelling as in [16], 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 [18].

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



(a) (b)

(c)

Fig. 2: Elaboration through the method in [18] 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.

obtained with standard imaging and with specialized imaging.

A. Some experiments on a selected page of the Poem

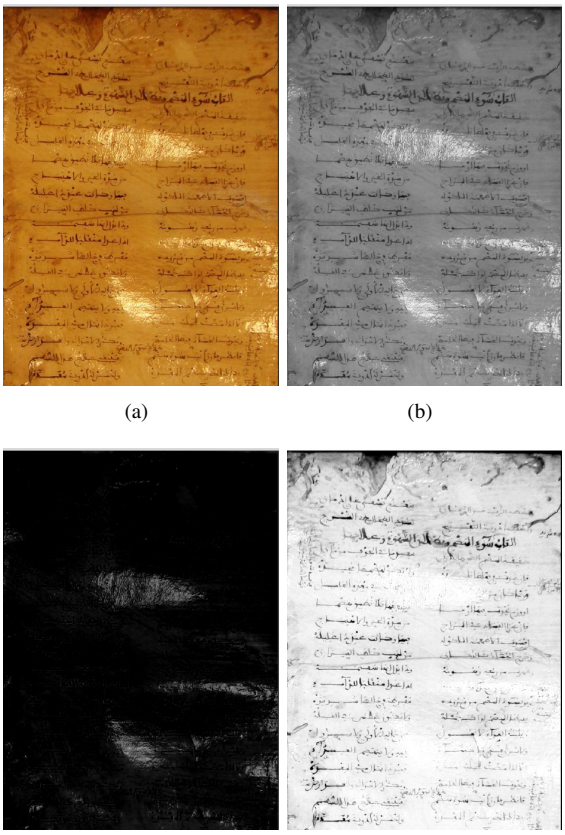
In this Section we presents some experiments of digitization and processing of the *62 verso* page, which is part of the Chapter 17 of the Section 3 of the Poem, and discusses *"the indigestion and lack of appetite"*.

We started by testing ICA directly on the educational RGB image available so far for this page. In this case, diversity of acquisition does not refer to different light polarizations, but to acquisitions in different wavelengths.

Figures 3(a) and 3(b) show the original images, in color and in grayscale, respectively. It is apparent that the strong reflection affecting this non-specialized digitization totally masks a good deal of the written content. Figures 3(c) and 3(d) show two components extracted through ICA. These well illustrate the effect of separation of text and reflection that we expect by ICA. However, the strong reflection prevented even a partial recovery of the masked text.

We then tested on the same image the algorithm proposed in [18], which could produce results as those shown in Figure 2. Unfortunately, in the digital educational version of the page, the reflected light completely masks the text. Hence, even that more sophisticated method was not effective, producing results very similar to those obtained with the simpler application of ICA (see Figure 3).

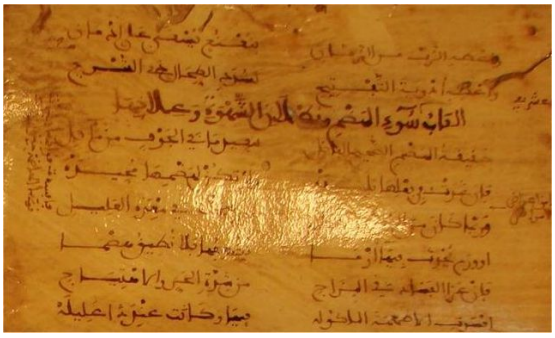
Hence, at present, we limited ourselves to perform accurate digitization only on those parts where the educational version was completely masked by the reflection, or the low resolution made impossible reading the text. Figure 4 shows a portion of that page that has interesting characteristics such as marginalia and scribe annotations.



(a) (b)

(c) (d)

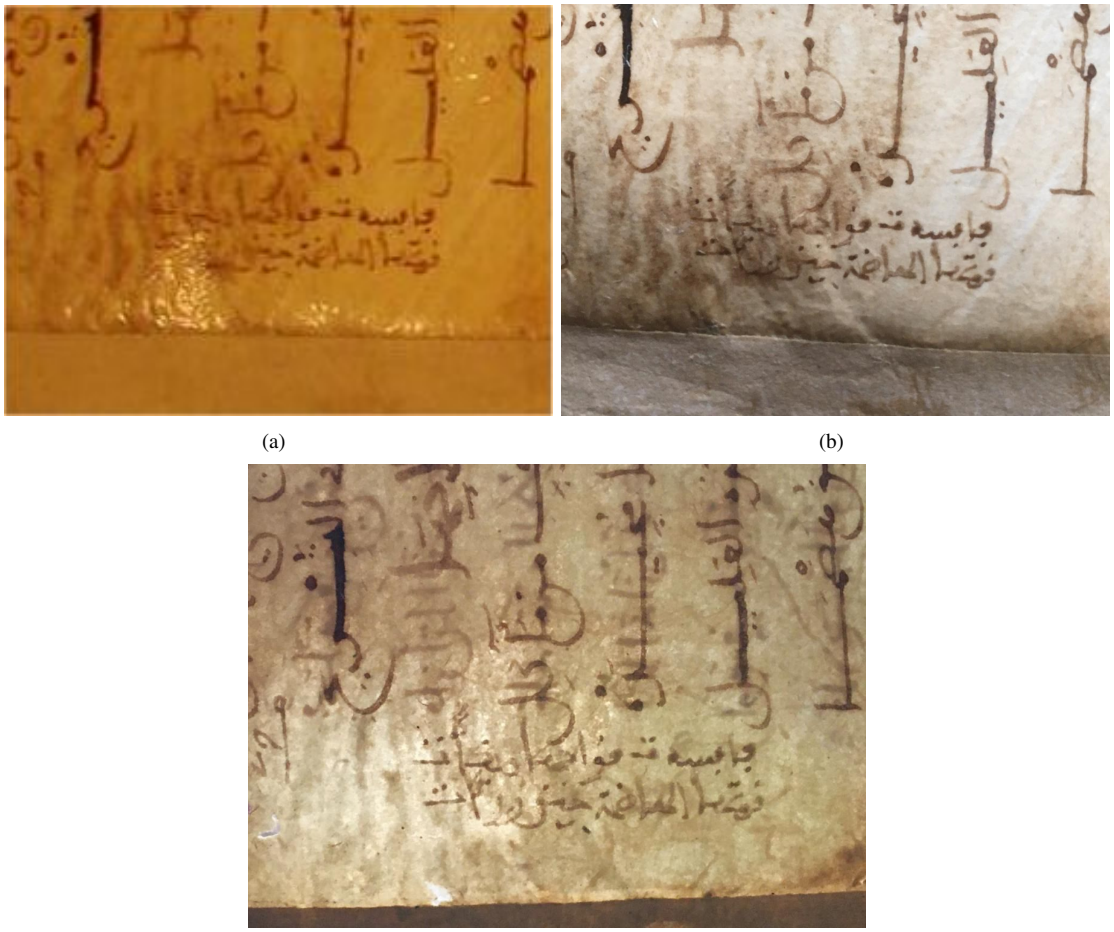
Fig. 3: Elaboration through ICA of the chosen page of the Poem: (a) the original RGB image; (b) the original grayscale image: (c) one ICA component showing the map of the reflected light; (d) another ICA component showing the text free of reflections. The dissatisfaction of the results is also partially due to the low resolution of the acquisition and the pdf compression of the educational images used.



(a)

Fig. 4: A portion of Figure 3 (a). Note the vertical marginalia on the left part, and the scribe annotations on the right margin.

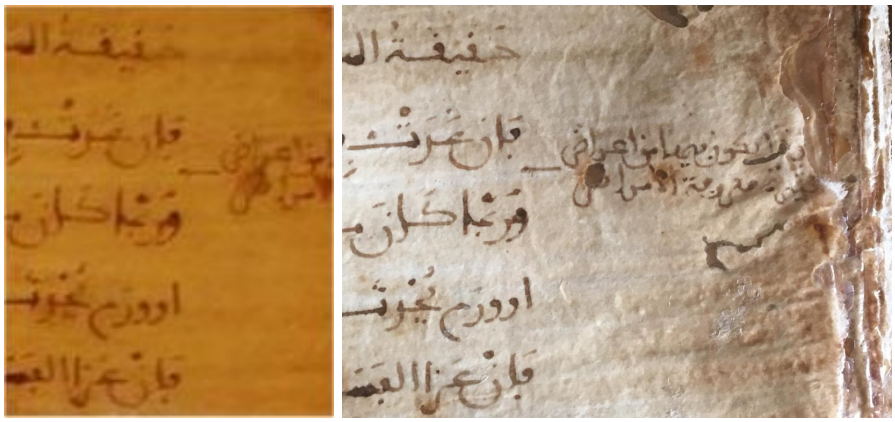
For the areas that presented reflection, or for fragments whose resolution was too low to permit readability, we repeated digitization locally, by accurately choosing the illumination setup and by using a more performing and high resolution camera. Figure 5 shows two different versions of



(a) (b)

(c)

Fig. 5: Variability of the obtained image when changing the illumination setup - a detail in the left part of the portion shown in Figure 4 (rotated): (a) the educational version, where the marginalia are unreadable due to reflection; (b) a specialized acquisition that avoids reflection and permits a good readability of the marginalia; (c) another acquisition obtained with yet another illumination setup - notably, in this case, extra characters shine through the back page.



(a) (b)

Fig. 6: Acquisition at a higher resolution of a scribe annotation in the right margin of the page (b), to be compared with the online educational version (a).

good quality digital acquisitions of a marginalia detail in the left part of the page, originally masked by the reflection. Figure 6 shows instead the acquisition at higher resolution of a scribe annotation in the right part of the page.

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 typical purposes of image processing is to improve legibility - understood as the ease of the reading, transcription, and linguistic analysis of the text conveyed by the manuscript.

In order to scholarly study and edit primary sources, a computational system has to provide effective support to encode and process 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.

As regards the software architecture, we are developing a home-designed platform made up of extensible and 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.

Thereby, the edition of the historical manuscripts is based on state-of-the-art methods in digital scholarly editing [23], and on a progressive publication process enabled by an interactive open source web platform. The under-construction platform is going to provide:

- advanced management functionality (zoom, rotation, 3D, YUV) for high-resolution facsimile reproductions (300 to 600 DPI);
- indexing and search functionality supported by gazettes and lists of named entities (person names, place names, analytical and iconographic indexes);
- advanced textual query functionality based on orthographic and linguistic features [19] [20].

In addition, the system has to handle:

- the collection of IIIF-compliant images;
- the transcription of textual content by means of collaborative editor tools;
- the text inquiry by means of intelligent search and indexing features;
- the encoding of the primary source with a special attention to the structure and the content of the original document.

B. Scholarly Text Encoding

Text - especially literary and/or historical one - is much more than a mere sequence of ordered characters. Indeed, textual 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 need to take into account such a multidimensional nature of their primary sources (e.g. physical, logical, historical, linguistic, communicative, etc.), and therefore their formal representation seeks to make explicit, machine readable and machine actionable all these aspects [29].

A shared model to scholarly encode texts and to digitally record linguistic analyses is the schema defined by the Text Encoding Initiative (TEI) [24]. [4] The TEI project is well organized in a modular framework, allowing effective customization and extensions to adapt it to scholar's requirements. 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 the needs textual scholars can exhibit. Furthermore, different software tools have been developed during the last decades to process and visualize documents encoded in TEI-XML format [30]–[34].

Figure 7 shows the main structure of the TEI document which represents the relevant divisions we defined for the scholarly encoding of the Poem. Such a digital representation[5] provides the TEI conforming document for our case study, showing the two principal document blocks, namely 1) metadata (`<teiHeader>`) and 2) text structure `<text>`. The main hierarchical relations we have encoded are the following:

- `<div1>`, which represents the original sections, i.e. the *maqālah* in the Poem. The original source includes seven sections, each of them discusses a part of the human body, the symptoms of the diseases that can be observed and, in the end, the treatments. Thus, we have seven `<div1>` elements throughout the XML document. The `div1` element brings properties as XML attributes that characterize some traits of the division. For example, we encode:
  - the type of the division: the `@type` attribute;
  - the identifier of the division: the `@xml:id` attribute;
  - the named order of the division: the `@n` attribute.

  Other elements, nested within the section division, encode the long title and the abbreviated one (`head`, `title`, `abbr` elements).
- `<div2>`, which represents the original chapters, i.e. the *bāb* in the Poem. Each section is divided into several chapters. For example, Figure 7 illustrates the chapter number 17 of the Section number 3, which begins on 62v page and discusses *the indigestion and lack of appetite* in 15 verses. The therapies are announced via a sub-title that has been marked through the `<fw>` element.
- `<l>`, which represents the level of verse lines for the Poem. In Arabic poetry, a line of verse almost invariably consists of two hemstitches. So, the `<l>` element is composed of two general purpose segmentation fragments (`<seg>` elements), which have been adopted to encode the two hemistiches (right and left, respectively).

Figure 8 shows various types of TEI elements to describe an example of the encoded fragment of the Poem:

[4]The TEI method is currently the de-facto standard to encode historical manuscripts assuming a philological perspective.

[5]The XML-TEI schema presented in figure 7 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).

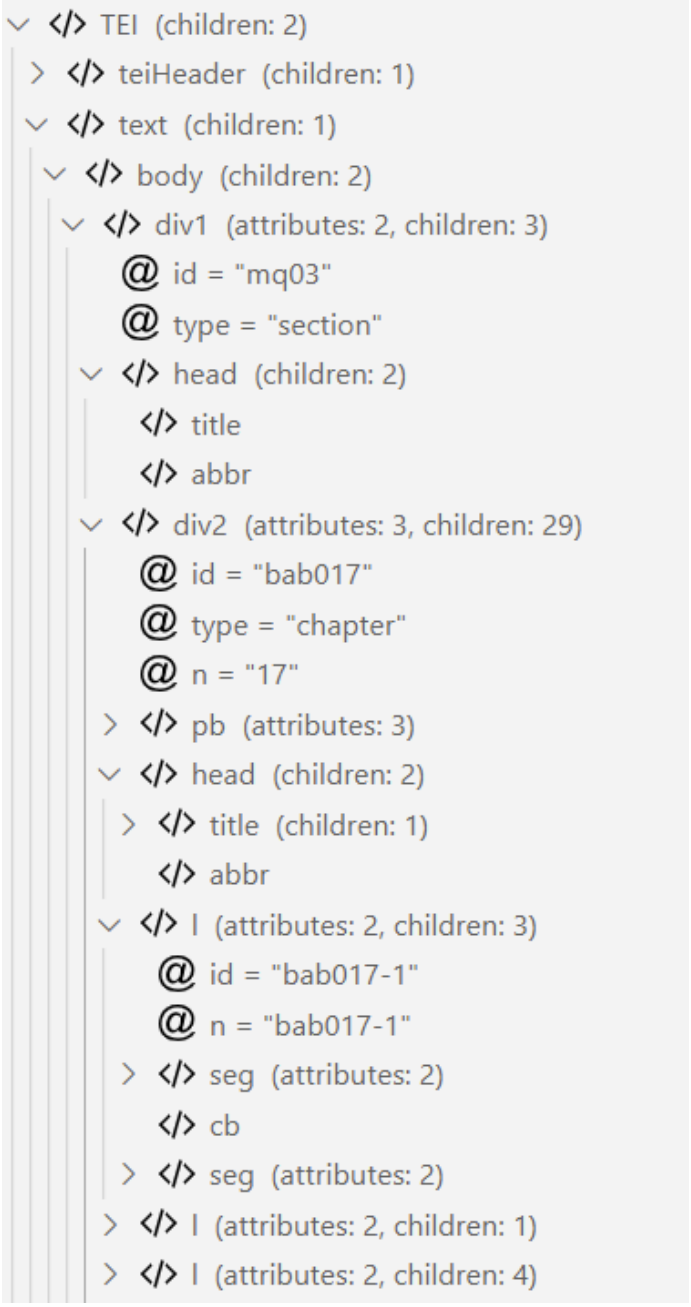


Fig. 7: The hierarchical structure of the Poem

1) verse lines, by using the `<l>` element;
2) hemistiches, represented by the `<seg>` element (right and left, respectively);
3) marginalia and addition annotations, by using `<add>` element (right and left of the page);
4) difficult readings, represented by the `<unclear>` element, which also records the responsibility of the text interpretation (i.e. who made the reading).

C. Linguistic Analysis

Linguistic analysis is an advanced representation of the content information conveyed by textual documents. Such an outcome is the result of processing the encoded text

```
<l xml:id="bab017-3" n="bab017-3">
  <seg type="h-bab017-3-hemistich" xml:id="bab017-3-right">
    فإن غدت في فعلها بليدة
  </seg>
  <cb/>
  <seg xml:id="h-bab017-3-hemistich" n="bab017-3-left">
    ولم تكن لهضمها مجيدة
  </seg>
</l>
<l xml:id="bab017-4" n="bab017-4">
  <add place="margin-right" rend="interlinear-dash">
    <seg type="h-bab017-4-hemistich" xml:id="bab017-4-right">
      بأن يكون فيها من أعراض
    </seg>
    <lb />
    <seg type="h-bab017-4-hemistich" xml:id="bab017-4-left">
      كثيرة معروفة الأمراض
    </seg>
  </add>
</l>
```

Fig. 8: Example of encoding for a fragment of the Poem

at different levels of complexity: morphological analysis, syntactical analysis, and semantic interpretation.

With specific reference to our case study, the process of lemmatization[6] requires, first of all, the vocalization of words, as the written Arabic does not contain vowels. This step is very demanding and entails scholar's responsibility, due to the "interpretation" and understanding of the text [21].

In order to encode the linguistic analysis of the Poem, we used the CoNLL-U format[7] where linguistic annotations are represented in tabular form through simple plain text files [28].

Figure 9 illustrates part of the poem in CoNLL-U format.

The text is subdivided using three types of lines:

- word lines which register the annotation of a word/token by means of five fields represented by single tab characters (see below);
- one blank line which marks hemistich boundaries;
- two blank lines which mark verse boundaries.

Word lines contain the following fields:

1) ID: words indexed with the identifiers taking into account the physical structure of the Poem. For example, the `ID=S3.C17-therapy.T1` corresponds to the first token of the Title ($T1$) of the Therapy, within the the Chapter 17 ($C17$) in the Section 3 ($S3$).
   When the token is a "multi-word", it is divided into sub-tokens that inherit the same ID adding a sequential number as a suffix.

[6]We call lemmatization the process of labelling a word with its basic written form or with its entry in authoritative dictionaries or lexicons.
[7]https://universaldependencies.org/format.html

```
# global.columns = ID FORM LEMMA POS FEATS MISC
# sent_id = Urjouwzah_Poem
# text = الأرْجُوزَةُ الطِّبِّيَّةُ لِابْنِ طُفَيْل
S3.C17-therapy.T1   عِلَاج عِلَاجُ NOUN    Animacy=Nhum|Case=Nom|Definite=Cons|Gender=Masc|Number=Sing
S3.C17-therapy.T2   سُوء سُوءِ NOUN    Animacy=Nhum|Case=Gen|Definite=Cons|Gender=Masc|Number=Sing
S3.C17-therapy.T3   الـهَضْم   هَضْم NOUN    Animacy=Nhum|Case=Gen|Definite=Def|Gender=Masc|Number=Sing


S3.C17-therapy-line1.E1.T1  فَـانْظُرْ  _
S3.C17-therapy-line1.E1.T1.1    فَ   فَ   PART    _
S3.C17-therapy-line1.E1.T1.2    انْظُرْ    نَظَرَ VERB    Gender=Masc|Mood=Imp|Number=Sing|Person=2
S3.C17-therapy-line1.E1.T2  فَـإِنْ _   _   _
S3.C17-therapy-line1.E1.T2.1    فَ   فَ   PART    _
S3.C17-therapy-line1.E1.T2.2    إِنْ   إِنْ   ADP _
S3.C17-therapy-line1.E1.T3  رَأَيْتَ  _
S3.C17-therapy-line1.E1.T3.1    رَأَى رَأَيْ VERB    Aspect=Perf|Gender=Masc|Number=Sing|Person=2|Tense=Past|Voice=Act
S3.C17-therapy-line1.E1.T3.2    تَ   _   PRON    Case=Nom|Number=Sing|Person=2
S3.C17-therapy-line1.E1.T4  سُوء سُوءَ NOUN    Animacy=Nhum|Case=Acc|Definite=Cons|Gender=Fem|Number=Sing
S3.C17-therapy-line1.E1.T5  هَضْم هَضْم NOUN    Animacy=Nhum|Case=Gen|Definite=Ind|Gender=Masc|Number=Sing

S3.C17-therapy-line1.E2.T1  فيـنبغي  _   VERB    Aspect=Perf|Gender=Fem|Number=Sing|Person=3|Tense=Past|Voice=Act
S3.C17-therapy-line1.E2.T1.1    فَ   فَ   PART    _
S3.C17-therapy-line1.E2.T1.2    يـنبغي   إِنْبَغَى   VERB    Aspect=Perf|Gender=Fem|Number=Sing|Person=3|Tense=Past|Voice=Act
S3.C17-therapy-line1.E2.T2  عِلَاجَ عِلَاجَ NOUN    Case=Acc|Definite=Cons|Gender=Fem
S3.C17-therapy-line1.E2.T3  هَذَا هَذَا PRON    Case=Gen|Gender=Masc|Number=Sing
S3.C17-therapy-line1.E2.T4  الـسَّقْم  _   NOUN    Animacy=Nhum|Case=Gen|Definite=Def|Gender=Masc|Number=Sing
S3.C17-therapy-line1.E2.T4.1    ال   _   NOUN    Animacy=Nhum|Case=Gen|Definite=Def|Gender=Masc|Number=Sing
S3.C17-therapy-line1.E2.T4.2    سَقْم سَقْم NOUN    Animacy=Nhum|Case=Gen|Definite=Def|Gender=Masc|Number=Sing
```

Fig. 9: Example of linguistic analysis.

For example, the word فَانْظُر (*fa=nẓur*[8] 'so look')[9] corresponds to the `first token` ($T1$) of the `right hemistich` ($E1$) of the `first verse` (*line*1) of the therapy. It is located within the `Chapter 17` ($C17$) in the `Section 3` ($S3$). So, it is identified by `ID=S3.C17-therapy-line1.E1.T1` and it consists of two sub-tokens:

- the first sub-token has the identifier `ID=S3.C17-therapy-line1.E1.T1.1` and corresponds to the conjunction فَ (*fa*, "so")
- the second sub-token انْظُر (*nẓur*) has `ID=S3.C17-therapy-line1.E1.T1.2` and corresponds to the imperative form of the verb نَظَرَ (*naẓara* "to see").

2) FORM: contains the word form or punctuation symbol; For example, the word *fa=nẓur* in the previous example;

3) LEMMA: contains the canonical form of the lexical entry. The token *fa=nẓur* is linked to the lemma *naẓara*;

4) UPOS: contains the part-of-speech tag of the word belonging to the tagset of the universal dependency; grammar;[10]

5) FEATS: contains the list of morphological features belonging to the universal feature inventory. For example, the word form *fa=nẓur* is the form of the second person to the imperative, *Gender=Masc—Mood=Imp—Number=Sing—Person=2*

In a second phase, the linguistic analyses are linked to the TEI-XML document by means of the `@key` attribute of the `w` element, as shown in Figure 10.

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:

- the relevant concepts, e.g. the anatomic structures, but also the concepts related to the diagnosis, the prognosis, etc.;
- the semantic relationships among the terms;
- the English term corresponding to which the domain term is linked

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 is سُوء (*sū'* "evil, ill") and it corresponds to the prefixes *'in-'* or *'mis-'*;
- the second word is هَضْم (*haḍm*), which means *'digestion'*.

Both of these two words have been linked to the term *"indigestion"* and annotated with the `@part` attribute.

The word ***sū'*** **"evil, ill"** with the identifier `id="S3.C17-therapy-line1.E1.T4` and the following word ***haḍm*** **"digestion"** with the identifier

[8]We use the Leipzig Glossing Rules, [27], for the "syntax" and "semantics" of interlinear glosses. Interlinear morpheme-by-morpheme glosses give information about the meanings and grammatical properties of individual words and parts of words. So, segmentable morphemes are separated by hyphens and clitic boundaries are marked by an equals sign.

[9]The Arabic word is followed by its transliteration in italic and its translations within quotation marks.

[10]https://universaldependencies.org/u/pos/index.html

```
<p xml:id="bab017-c1">
    <fw type="cure">عِلَاجُ سُوءِ الـهَضْم</fw>
</p>
<lg corresp="#c15-1">
    <l xml:id="bab017-c1-1" n="1">
        <seg xml:id="h-bab017-c1-1-left" n="bab017-c1-1-1-left">
            <w xml:id="S3.C17-therapy-line1.E1.T1" key="T1.1-T1.2">فَـانْظُرْ</w>
            <w xml:id="S3.C17-therapy-line1.E1.T2" key="T2.1-T2.2">فَـإِنْ</w>
            <w xml:id="S3.C17-therapy-line1.E1.T3" key="T3.1-T3.2">رَ أَيْتَ</w>
            <w xml:id="S3.C17-therapy-line1.E1.T4" key="T4" type="domain" lemmaRef="#indigestion" part="I">سُوءَ</w>
            <w xml:id="S3.C17-therapy-line1.E1.T5" key="T5" type="domain" lemmaRef="#indigestion" part="F">هَضْم</w>
        </seg>

        <seg xml:id="h-bab017-c1-1-right" n="bab017-c1-1-1-right">
            <w xml:id="S3.C17-therapy-line1.E2.T1" key="T1.1-T1.2">فَـبْتَغِي</w>
            <w xml:id="S3.C17-therapy-line1.E2.T2" key="T2" type="domain" lemma="عِلَاج" pos="Noun"
               lemmaRef="#cure">عِلَاجَ</w>
            <w xml:id="S3.C17-therapy-line1.E2.T3" key="T3" >هَذَا</w>
            <w xml:id="S3.C17-therapy-line1.E2.T4" key="T4.1-T4.2" type="domain" lemmaRef="#sickness">الـسَّقْم</w>
        </seg>
    </l>
```

Fig. 10: Example of encoding the therapy block of verses and the linguistic annotations along with single and polyrematic terms.

id="S3.C17-therapy-line1.E1.T5"[11] are part of the domain term "digestion" and have the @part attribute with value "I" (Initial) and "F" (Final) respectively:

```
<w
    xml:id="S3.C17-therapy-line1.E1.T4"
    key="T4" type="domain"
    lemmaRef="#indigestion"
    part="I">
    سُوء
</w>
<w
    xml:id="S3.C17-therapy-line1.E1.T5"
    key="T5" type="domain"
    lemmaRef="#indigestion"
    part="F">
    هَضم
</w>
```

IV. Conclusions

The text preservation and the digital restoration of ancient manuscripts are challenging tasks. Within this paper, we have depicted the architecture and the functionalities of a web-based system devoted to deal with the digital scholarly editing process and the digital availability of historical manuscripts. This large amount of documental heritage often is in degraded conditions due to the lamination process carried out on the manuscripts in the last decades.

The ongoing process that we are developing encompasses at least six steps:

1) digital acquisition of the primary source;
2) digital restoration of the facsimile scans;
3) transcription of the primary source;
4) encoding of the structures, contents and phenomena conveyed by the text;
5) performing linguistic analysis of the encoded text;
6) linking linguistic analyses to the original text.

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 physical support. Subsequent image processing techniques have been revised that can eliminate the residual reflection from the acquired images.

As far as the transcription and encoding phases are concerned, we used the Text Encoding Initiative guidelines (TEI) in order to scholarly prepare a digital representation of the primary source. TEI is an XML vocabulary devoted to suggest best practices and standard encodings of textual phenomena. We encode divisions, verses, hemistiches, marginal additions, tokens and other significant features. The linguistic analysis are made by using the CoNLL-U format, which is an optimal method to facilitate the specialist work. After that, the analysis are liked to the text contents leveraging the TEI encoding expressiveness.

We have evaluated the feasibility of the proposed method on the chapter 17 in section 3 of the "Poem in Rajaz on medicine". This poem is unique and there is only one copy preserved by the Al Quaraouiyine Library located in Fez, Morocco. It is about a medicine poem, which describes the diseases symptoms and the relative cure. The diseases involve all organs from head to kidneys. The chapter 17 in section 3 concerns "indigestion, lack of appetite and care".

The few examples provided are encouraging us to carry on with this line of activity. Indeed, the improvement of the readability of the text by specialized digitization techniques and subsequent digital elaborations can significantly enhance and support the process of scholarly transcription, text encoding and linguistic analysis, performed through specialized semi-automatic computational tools.

[11] fourth and fifth tokens of the first line of the therapy, within the chapter 17 of the section 3.

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 Ṭufayl's ʾurguzah shows his educational side.

[3] M. Salih, Ibn Ṭufayl - qaḍāayāa wa=mawāaqif. dāar ar-raŝīid li-n-naŜr. 1980, pp.75-80

[4] M. Salih, Ibn Ṭufayl - qaḍāyā wa-mawāqif. dār ar-raŝīd li-n-naŝr. 1980. p.75-80

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

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

[7] 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).

[8] 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.

[9] 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).

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

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

[12] 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).

[13] 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).

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

[15] 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).

[16] 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.

[17] 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.

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

[19] 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.

[20] A. M. Del Grosso and 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.

[21] 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).

[22] O. Nahli and 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.

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

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

[25] A. M. Del Grosso, D. F. Fihri, M. El Mohajir, O. Nahli, and A. Tonazzini, "Digital Safeguard of Laminated Historical Manuscripts: The Treatise "Poem in Rajaz on Medicine" as a Case Study", DPWH, Cist 2020, Agadir, Morocco, IEEE XPLORE, pp. 192-197, DOI: 1https://doi.org/10.1109/CIST49399.2021.9357192.

[26] A. Hyvarinen, J. Karhunen, and E. Oja, Independent Component Analysis (Wiley, New York, 2001).

[27] B. Comrie, M. Haspelmath, and B. Bickel, The Leipzig GlossingRules, "Conventions for Interlinear Morpheme-by-morphene Glosses". Max Planck Institute for Evolutionary Anthropology, 2008. Available: https://www.eva.mpg.de/lingua/resources/glossing-rules.php

[28] S. Buchholz and E. Marsi, "CoNLL-X shared task on Multilingual Dependency Parsing". Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), pages 149–164. New York City, 2006.

[29] Christof Schöch, José Calvo Tello, Ulrike Henny-Krahmer and Stefanie Popp, "The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in TEI XML", Journal of the Text Encoding Initiative [Online], Rolling Issue, Online since 14 August 2019, connection on 07 May 2021. URL: http://journals.openedition.org/jtei/2085; DOI: https://doi.org/10.4000/jtei.2085

[30] Heiden, Serge. 2010. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation - PACLIC24, edited by Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, 389–98. Sendai: Waseda University. https://www.aclweb.org/anthology/Y10-1044. Also available at https://halshs.archives-ouvertes.fr/halshs-00549764/en.

[31] Roberto Rosselli Del Turco, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti and Jacopo Pugliese, "Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions", Journal of the Text Encoding Initiative [Online], Issue 8 — December 2014 - December 2015, Online since 29 December 2014, connection on 07 May 2021. URL: http://journals.openedition.org/jtei/1077; DOI: https://doi.org/10.4000/jtei.1077

[32] Julia Flanders and Scott Hamlin, "TAPAS: Building a TEI Publishing and Repository Service", Journal of the Text Encoding Initiative [Online], Issue 5 — June 2013, Online since 24 June 2013, connection on 07 May 2021. URL: http://journals.openedition.org/jtei/788; DOI: https://doi.org/10.4000/jtei.788

[33] TeiPublisher: https://teipublisher.com/index.html

[34] eXtensible Text Framework (XTF):https://xtf.cdlib.org/