Methods of the Arabic Manuscripts Digitization[1]

Prof. Oleg Redkin, Dr. Olga Bernikova
Department of Asian and African Studies, Laboratory for Analysis and Modeling of the Social Processes,
St. Petersburg State University, St. Petersburg, Russia

Abstract *The mediaeval Arabic manuscripts are not only valuable artifacts but they also represent one of the major sources of scholar information in the field of Oriental Studies. This paper discusses the methods of Arabic Manuscripts Digitization. Over the last fifteen years a lot of Arabic manuscript digitization projects have been carried out. Digital collections of the manuscripts based on Arabic script are represented in the collections of libraries worldwide, including on-line databases. Nevertheless, these collections are restricted by their functionality: technology of metadata integration relies on the human made characteristics. While a possibility of automatic metadata introduction would facilitate the task of manuscript processing, at the same time it allows automatic quantitative and quality analysis of the manuscripts. This paper analyses different methods for classifying and retrieving historical Arabic handwritten documents and suggests the most efficient methods of their digitization.*

Keywords: manuscript, digitization, Arabic

1 Introduction

Mediaeval Arabic manuscripts are not only valuable artifacts but they represent one of the major sources of scholar information in the field of Oriental Studies as well Although Arabic manuscripts have always remained in the focus of the scholars' attention, for a long period of time the methods of their description and investigation have been almost unchanged and based not only on researcher's experience, qualification and knowledge, but on a subjunctive opinion as well.

The description of these manuscripts has a long history and, as a rule, includes a collection of data on the history of their origin, content and characteristics of the physical state of the document.

Recent decades have witnessed the spread of the digital processing, retrieval, storage and transmission of information which, in its turn, has allowed new methods of data processing in Arabic, and opened new opportunities for scholars. Thus the digitalization of Arabic handwriting heritage has completely revolutionized this process and provides creation of electronic on-line catalogs, the digitization of the scanned images and, to some extent, optical character recognition (OCR).

2 The term “digitization”

In the historical perspective “digitizing of a document” meant a creative surrogate, an alternative carrier intended to be preserved [2]. Today there are several different interpretations of the term. Simplified understanding of the first approach is digitizing as making images: computer processing of Arabic manuscripts limited to their scanning and recording received in *.bmp, *.jpg, *.ipeg or other types of files on the media or posting them on sites of other academic institutions. The second approach lies in the field of text recognition, i.e. digitizing that includes scanning and optical character recognition as a minimum. This solution is quite difficult in case of Arabic manuscripts. There is another interpretation of the term “digitizing” which we refer to a historical document. Digitizing a huge amount of manuscripts requires a sophisticated information system that established relations between data (digital images) and metadata.

Metadata is a “structured piece of information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" [3]. As a minimum, metadata should conform to the Shareable Metadata Guidelines for libraries. Digitization enhances access to the artifact as its image can be seen on the web by users all over the world. Besides, it can be sent for offline viewing using a higher resolution of uncompressed master file.

3 Previous experiences

Over the last fifteen years a lot of Arabic manuscripts digitization projects were carried out. Digital collection of manuscripts based on Arabic script is represented in the collections of libraries worldwide, including on-line databases [4]. For example, Princeton University Library and the Free University, Berlin, in conjunction with the Imam Zayd ibn Ali Cultural Foundation (IZbACF) in Sanaa, Yemen [5] implemented the collection that is a part of the Princeton Digital Library of Islamic Manuscripts [6].

[1] The authors acknowledge Saint-Petersburg State University for a research grant 2.37.175.2014.

One of the best manuscripts database is presented by the Welcome Arabic Manuscript Cataloguing Partnership (WAMCP) that combines the efforts of the Welcome Library (London), Bibliotheca Alexandrina (Alexandria, Egypt) and Department of Digital Humanities (King's College, London) [7]. The interface of the website is user-friendly. Both manuscript and metadata can be compared side by side. All indexed fields are searchable. A positive issue is the additional possibility to search in full text as well as in one of the following fields only: incipit, colophon of the manuscript, colophon of work, table of contents, notes and provenance. At the same time it is worth noticing the lack of detailed information about the amount of the digitized manuscripts and their most common topics which are primarily related to medicine.

Significantly, the system allows searching via the full text search facility. We can hardly check the correctness of this kind of searching technology. Today automatic recognition of handwritten words remains a challenging task. To a certain extent it is caused by the peculiarities of the Arabic text (it was described in details in our previous works, see [8] and [9]). The problem of recognition of handwritten documents, especially manuscripts, which include the individual characteristics of the authors' handwriting, is even more complex, not to mention the extra noises – notes of the scribes, defects of the written material, and lacunae and gaps in the text, notes and additions to the original text. All this makes the correct identification of the Arabic written texts extremely difficult. One of the experiments in this field was carried out in terms of the developing a new database with handwritten Arabic town/village names [10]. It proves the fact that error rate of handwritten recognition systems are much better for applications with a restricted lexicon of words. That is why the searching mechanism presented in Welcome Arabic Manuscripts Online requires clarification: is it really based on the full-text manuscript search or it relies on restricted parts of the retyped manuscript? We assume that the searching mechanism relies only on those parts of manuscripts that have printed form. Our methods of manuscript digitizing and classification strongly differ from the technologies used in WAMCP, as they are concentrated on different tasks. WAMCP used manual processing of the metadata manuscripts while our experiment is targeted at the basic automatic manuscript processing. Our search for similar available solutions confirmed the uniqueness of statements of problems. Nevertheless, a group of scholars from the Cairo University investigated similar issues. They presented results for historical document classification of old Arabic manuscripts using texture analysis and a segmentation free approach. The main purpose of their project was "to discriminate between historical documents of different writing styles to three different ages: Contemporary (Modern) Age, Ottoman Age and Mamluk Age" [11].

4 Specificity of Arabic Manuscript Digitization

Currently most of the manuscript centers and libraries focus on the creation of a digital copy of the manuscripts which in fact is the formation of databases either within the library network, or on-line supplied with search engines, rather than processing of the special features of a certain manuscript. The related and additional information is downloaded manually, and criteria of its selection depend on the operator's subjunctive opinion.

For unification of the manuscript digitization the UMass Amherst Libraries Guidelines for Digitization were developed by members of the Digital Creation and Preservation Group for the application of all library digitization projects. These guidelines are designed to provide digital project managers with a set of minimum specifications for preservation-quality digitization of printed text, manuscripts, photographs, slides, rare books, sheet music, graphic arts, and maps. They provide a baseline for creating digital images that are of sufficient quality for long-term preservation [12].

Traditionally so called "subject metadata" include manuscript number, title, author, origin, organization, commentary, commentator, language of material, script, complete/incomplete, folio number, script, subject, bundle number, folio number, pages, missing portion, illustrations, condition, catalogue source, remarks, manuscript date, manuscript length, manuscript width. As well as the information about its origin, incipit, explicit, physical details, which include information about the script, color of ink, quality and completeness of the manuscript itself, number of lines on the page, columns, binding, loose quires etc.

The majority of the existing databases rely upon manual processing of the manuscript with its following integration into the software environment. At the same time existing technological solutions allow for the automatic description of a number of metadata characteristics. As a result there are two methods of manuscript analysis: the first of them is based upon the results of subjective perception and the second one is based on objective quantitative and qualitative indicators, which are presented in digital form.

The objective data usually refer only to the number and size of the pages, type of the binding, paper quality, number of lines per page, etc. More detailed characteristics, for example those associated with the so-called "rhythm" of the handwriting, with few exceptions still remain outside the view of researchers.

Meanwhile, the more detailed analysis of the formal data may provide additional information that can draw conclusions about the time and the place of the manuscript origin.

On the contrary, the computer analysis of the text may enable automatic determination of the whole range of characteristics such as:

- the type of the script;

- proportions between the size of the fields and text spacing between words and long words;

- the ratio between the height and width of the characters, diacritics location, tilt of the handwriting;

- the width of the lines of the characters can determine the degree of pressure, and hence the written tool type;

- the presence of different types of handwriting in one manuscript;

- the ratio between the text and the gaps;

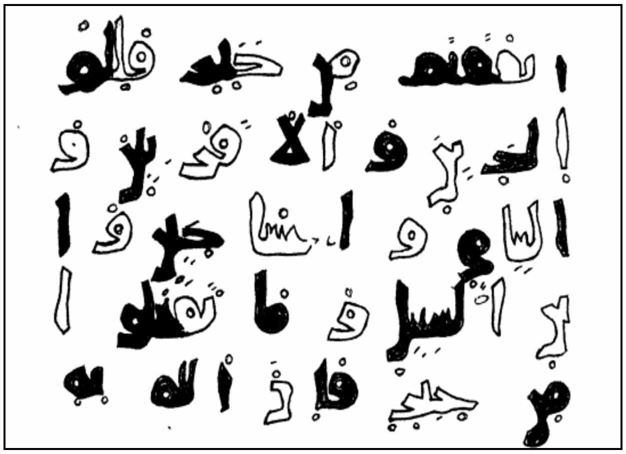
- to identify particular color palette, illustrations, inventory marks and seals, inscriptions on the sidelines, the color characteristics of ink, the use of color in the text of the manuscript;

- to define ratio between the parameters of the text and the features of its and orientation of its fragments, the shape and the type of the manuscript layout.

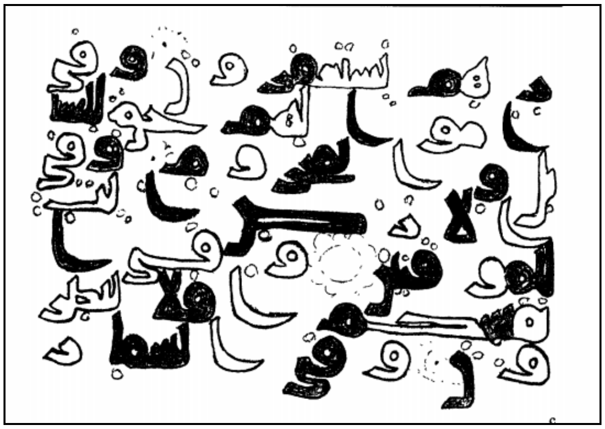
Due to the objective assessment of each manuscript the text may be correlated with a particular school of manuscript, in fact such a correlation is possible when the information is available on the manuscript scribes.

In this case the traditional methods of the classification of the script are insufficient since there are also "hybrid" versions of the characters (eg. Maghreb typeface of) ـفfā ') / ق (ķāf)).

It is the computer analysis which can provide scale and objective systematization of ideas about spellings lower and upper detail view, closed) ـمmīm) and open) ـبbā') inner elements of the characters, diacritics, vowels, ligatures.



Picture 1a. Types of manuscripts which are found different in diagonal effects [13].



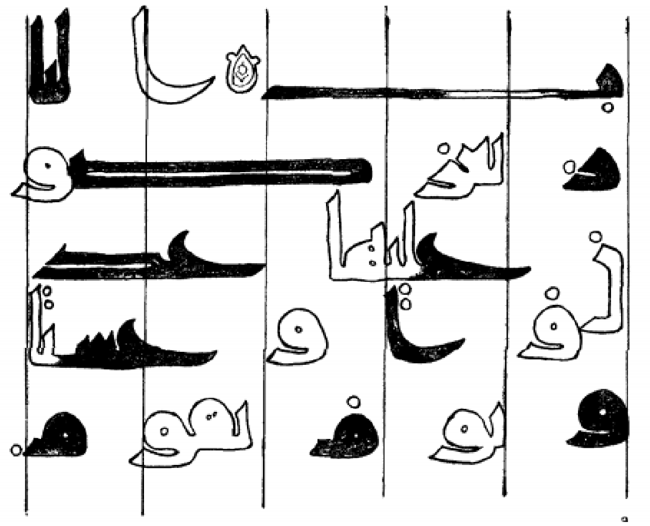
Picture 1b. Types of manuscripts which are found different in diagonal effects [13].

The traditional classification of Arabic handwriting scripts (naskh, kufi, maghribi, nasta'liq etc.) which is based on subjective visual analysis and evaluation remains unchanged, and it is rather arbitrary (‘number of points along the lines’), and does not reflect the whole reality of Arabic tradition of handwriting.

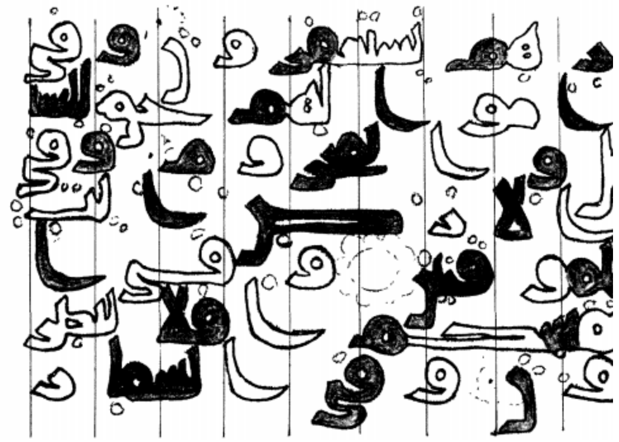
Computer analysis would allow going far beyond the approximate classification of Arabic handwriting. It will help to classify handwritten texts based on objective (digital) characteristics, for example, based on such indicators as the differences in the degree of roundness of letters in Naskh and Kufic texts. As a result, the whole range of the script variants may be divided into several groups (clusters) depending on the type of characters, their density, etc.

It is known that texts which belong to different handwritten schools have different writing of certain letters, as well as the differences in the location and orientation of the fragments of the text, text density (space occupied by individual letters, the number of letters falling on the page), variety of extra elements, such as illumination, colors. Thus, the task is to identify certain patterns and proportions of the text rather than to implement linear and vertical segmentation, as it is customary in the OCR.

Diachronically the basic parameters of the Arabic manuscripts were developing (i.e. size of the pages, type of binding, ink color, its type) depending on the manufacturer, written tools, local technologies and existing manuscript tradition (line width, spacing, types of word wraps, peculiarities of graphics, text illuminations etc.). Since all these parameters are multifaceted and include a lot of data, they cannot be systemized and described as a whole without the application of computer technologies.



Picture 2a. Division of the text into different segments [13].



Picture 2b. Division of the text into different segments [13].

5 Databases and digital ‘passport’ of the manuscript

The spectrum and combination of digital indicators of this kind are unique for each manuscript, and may be considered as manuscript digital ID.

Each manuscript has a unique set of special features and characteristics (types of script, variations of color, filigrees, chainlines, various types of paper, etc.) which may be compared with fingerprints, and which may be processed with the help of computer technologies. Thus, digital processing can determine the exact number of lines in the text, the angle at which the text is placed between the grid lines (laid paper / chainlines) on a specific page, as well as throughout the manuscript as a whole. In this case the task is to determine which features are relevant to a certain manuscript and, finally, to build hierarchical system of these features. The data obtained will allow formation of databases and developing advanced search engines.

Characteristics of the manuscript can be divided into those which ‘lie on a surface’ (explicit), and those that require further analysis (implicit).

The language of the manuscript, handwritten text color, manuscript size, number of pages, paper color, text orientation (landscape / non-album) etc. are among these explicit characteristics. Among the implicit characteristics are the number of characters, the ‘inner rhythmic of the text’, location and orientation of its elements, use of words and expressions, etc.

Digital 'passport' forms a manuscript's feature vector which may be interpreted as a point in a metric space. Having a set of feature vectors it is possible to perform a clustering (grouping) of manuscripts. On the one hand manuscript clustering allows to build a classifier to automatically determine the writing style, era and even the possible authorship. On the other hand it allows to extract an essential data for analysis of relationships between manuscripts and for tracking the dynamics of writing style evolution. It is known that clustering methods on the basis of randomized learning theory [14, 15] provide an appropriate results for different kind of data sets with noises and uncertainties. Such useful properties seem to be very promising for Arabic manuscripts processing.

6 Conclusion

The set of indicators specific to each individual manuscript is unique, and allows us to classify it and assign to a specific point in time and the spatial coordinates. Creating of such digital manuscript “passport” will allow its comparative analysis in comparison with other manuscripts, to implement their classification, to draw conclusions about its authorship, as well as to define whether it is an autographed manuscript or a copy, attribute it to a certain handwritten school or copyist, and also to help in its dating. Such kind of digital "passport" will facilitate the process of cataloging manuscripts, preparing them for further scholar investigation.

7 References

[1] Gacek A. “Some Remarks on the Cataloguing of Arabic Manuscripts”; British Society for Middle Eastern Studies, Vol. 10, No. 2 (1983). Pp. 173-179.

[2] Zdeněk U. “Digitization is not only making images: manuscript studies and digital processing of manuscripts”; Knygotyra. 2008. 51. Pp. 148-160.

[3] Understanding Metadata. NISO, 2004. Pp. 1-20.

[4]http://guides.lib.umich.edu/islamicmsstudies/onlinecollections

[5] http://wamcp.bibalex.org/about-us, 03/03/2014/

[6] http://pudl.princeton.edu/objects/9s1616928

[7] http://pudl.princeton.edu/collections/pudl0032

[8] Redkin O.I., Bernikova O.A. “Problems of the Arabic OCR: New Attitudes”; Proceedings of the 2013 International Conference on Artificial Intelligence. Las Vegas, USA, 2013. Pp. 777-782.

[9] Redkin O.I., Bernikova O.A. “On the Optical Character Recognition and Machine Translation Technology in Arabic”; Proceedings of the 2012 International Conference on Artificial Intelligence. Las Vegas, USA, 2011. Pp. 861-867.

[10] Pechwitz M., S. Snoussi Maddouri, Märgner V., Ellouze N., Amiri H., “IFN/ENIT-Database of Handwritten Arabic Words”; 7th Colloque International Francophone sur l'Ecrit et le Document , CIFED 2002, Oct. 21-23, 2002, Hammamet, Tunis, (2002). Pp. 1-8.

[11] Ahmad M. Abd Al-Aziz, M.Gheith, Ayman F. Sayed Recognition for old Arabic manuscripts using spatial gray level dependence (SGLD); Egyptian Informatics Journal, 2011, 12. Pp. 37-43.

[12] Banach M., Shelburne B., Shepherd K., Rubenstein A. Guidelines for Digitization, Digital Creation and Preservation Working Group. 2010-2011.

[13] Polosin V. “Manuscripts of Ibn Muqla’s Calligraphical School (the problem of their identification); Oriental Written Sources, The Institute of Oriental Manuscripts. 2004. Pp. 160-177.

[14] Avros R., Granichin O., Shalymov D., Volkovich Z., Weber G. “Randomized Algorithm of Finding the True Number of Clusters Based on Chebychev Polynomial Approximation”, Data Mining Foundations and Intelligent Paradigms, Vol. 1: Clustering, Association and Classification. Springer-Verlag, 2012.

[15] Granichin O., Shalymov D., Avros R., Volkovich Z., “A randomized algorithm for estimating the number of clusters”, Automation and remote control, 2011, V. 72(4) Pp. 754-765.