Al-Ādāb Magazine Archives: Digitization, Preservation and Access:

A Joint venture between Al-Ādāb Magazine and University Libraries of American University of Beirut

Basma Chebani, Head of Cataloguing ad Metadata Services Department, University Libraries, American University of Beirut

Elie Kahale, Head of Digital Initiatives and Imaging Department, University Libraries, American University of Beirut

Abstract

This paper will describe the main activities used to digitize Al-Ādāb Magazine for preservation and access purposes along the difficulties faced in digitization projects especially when dealing with Arabic language.

The main objective of the archiving project is to preserve Al-Ādāb Magazine, a journal that was published in print since 1953 till 2012, and to produce a searchable full text database of Al-Ādāb Magazine. Users will eventually be able to browse and access the metadata and the full text of the journal online at the volume, issue and article levels. Full text and index search by author, title and subject will be available using simple and advanced search techniques. Full text articles will be available for downloading. The searching on the metadata (author, title and subjects) and the full text will be available in an open source content management system.

Starting with Metadata creation, the Library Information System (Millennium) was used for the creation of the bibliographic records based on MARC 21 format. As for the indexing, the Library adopted the Library of Congress Subject Headings translated into Arabic by AUB University Libraries (UL). The MARC format data output could be converted to any metadata schema such as Dublin Core (DC) in order to be ready for uploading into the open source Content Management System (CMS).

The Library adopts a long term digital preservation strategy for the digitization by using international standards for digital file format (Tiff 6.0) and according to OAIS data model for digital preservation and Dublin Core (DC) for descriptive metadata.

All digital images will be converted to full text digital objects by using OCR (Optical Character Recognition) software which transforms the text image into searchable Arabic text by comparing the form of the letters with a list of Arabic characters in different fonts stored in SAKHR OCR program "Arabic Reader" used during the OCR process.

The main challenge was dealing with the printed publications that continuously changed and evolved through decades to follow different trends at different levels ranging from printing style, fonts, content, political instability and censorship, and so on. This affected several activities such as scanning, OCRing, metadata creation and others. Different approaches were taken to deal with such scenarios according to the period in which the journal was published

between 1953 and 2012. The accuracy of old issues resulting from the OCR is less than the accuracy obtained by the OCR in recent issues mainly during the 21th century.

I- Introduction

In the last few decades, the drastic evolvement of technology led to major changes in the area of education, learning and research which affected fundamentally the services provided in colleges, universities and Libraries. This created new challenges and opportunities such as the emergence of digital libraries. In her book "Exploring Digital Libraries Foundations, practice, prospects" Karen Calhoun (2014) points out that "… the beginning of digital libraries in 1991, the year in which the National Science Foundation (NSF) in the US sponsored a series of workshops on how to make digital libraries a reality, not just a dream." Digital Libraries can have different scopes and are usually created to target specific audience and needs based on a unique or set of digital collections.

The Al-Ādāb Magazine Archives digitized collection is mainly concerned with digitized Arabic text material. As Abby Smith (2001) noted in her report "Strategies for Building Digitized Collections" that "... libraries usually identify two reasons for digitization: to preserve analog collections and to extend the reach of those collections". The process of digitization is part of a process that entails several activities such as selection and assessment, material preparation, metadata creation, image capture and image enhancements, OCRing in case of text material and uploading the digital content to a repository or/and to publish it to a mainly a web content management system. All this process is governed by intellectual assets copyrights, quality control and continuous evaluation for efficient work. Finally, a continuous collaboration between librarians and subject matter experts is needed to reach the scope of the digitized collections.

Based on an agreement between AUB and the Editor in Chief of "Al-Ādāb" Magazine who is also the owner Dr. Samah Idriss, AUB UL indexed and made available online the Al-Ādāb archives since its initial publishing in 1953 till 2012 a total of 60 years of printed archives. This started by digitizing old issues and extracting text to make it searchable. Then by preparing the metadata for its content at the articles level to allow researcher and scholar to use basic and advanced search to look for articles by issues, author, subject, keywords and dates in addition to search within the text with the ability to download the full article.

This is part of AUB University Libraries digitization initiatives that attempt to preserve National and AUB cultural heritage in addition to disseminate information and promote knowledge by allowing access to AUB Community, to scholars, to researchers and to the largest possible audience.

Why Al-Ādāb Magazine?

Al-Ādāb Magazine is a literary and cultural Journal established in 1953 by Dr. Souhail Idris and focused on movements in literature and culture in the Arab world. It included files in

political thought, poetry, novel, short stories, movies criticism, theater, and general culture. It was considered as Arab cultural platform, despite the Arab States censorship. Al-Ādāb ceased publishing at the end of year 2012. Then it started as an online Journal in 2015 where it is also publishing part of the Archives that was already digitized by AUB UL

In the following sections, there is an overview about the methodology that was applied during the digitization process by describing the main activities that were conducted and identifying the standards that were adopted. Then there is a listing of the main challenges that were faced during the digitization process at different levels with adopted solutions to overcome such obstacles.

II- Methodology

1. Image Capture and Enhancement

Before starting the digitization process, two copies of each issue were secured to ensure that the scanning process will not be affected by missing pages, tight biding or any problem that will forbid of having clear pages ready for the OCR process. The benchmark used for digital images consisted of 300 dpi black and white except for the covers they were scanned at 24-bit color. All master digital images were TIFF format as per "Technical Guidelines for Digitizing Cultural Heritage Materials (2010)" by FADGI recommendations. The master files were scanned using mainly Planetary Scanners, where by around 50,000 digital master files were created from the 60 scanned volumes and are backed up on the AUB UL servers.

After finalizing the scanning phase, derivative copies were created from the master files in order to enhance the digital images by removing black edges, applying de-speckle, de-skewing and removing unwanted noise and lines that might affect later on the OCR process.

The image capture and image processing workflow go through a quality control process to minimize possible errors because any mistake will be harder to fix at a later stage and will cause a huge overhead to fix it. For example, if a page was scanned twice or a missing page was discovered all succeeding page numbering will need to be fixed in all phases of the workflow.

2. Descriptive Metadata

Since we need to create analytical metadata for the digitized articles of Al-Ādāb Magazine at the article level, we opted to use the Cataloging module of Millennium Library System using MARC 21 format. This helped a lot in the consistency of entries for authors and subject authorities because we took advantage of the authority file of the library system for controlling the authors and subjects.

We used AACR2 for cataloging of articles and we used AUB authorities for names and the Library of Congress Subject Headings translated into Arabic for the indexing because it is much more accurate to use controlled vocabularies instead of using free keywords which might scatter the same concept under many terms or subjects. It might also use more than one form for the same author, because the magazine lasted for 60 years when the concepts and the terms were in constant change due to the emerging of new philosophical and literary schools from modernism and post modernism to structuralism and existentialism. The controlled vocabularies will certainly reduce the noise in searching. The subjects were enhanced by adding the date and by the use of qualifiers which determine the type of article under a limited number of types (Study, File, Review, Poetry, Novel. Criticism, Conference, etc.) in addition these categories will help in filtering the search results.

Once the metadata entry of articles was done, this metadata was exported in csv (comma separated values) format according to a simple mapping to Dublin Core elements to be uploaded later in XTF Content Management System (CMS). Dublin Core is the schema used for the metadata in (XTF).

3. Optical Character Recognition

Before uploading the digitized articles to the Content Management System there is a need to convert the scanned images of articles from digital images into text in order to enable the full text indexing by "Lucene" to be used later on by XTF search engine.

The Optical Character Recognition (OCR) is a technique for transforming the text from an image or a scanned document by a scanner or a digital camera, to a digital text that can be edited, formatted, searched and indexed. This technique spares the efforts of re-typing the texts in a text editor or word processor. Note that the term character is used instead of letter because character means letter, symbol, number, punctuation mark or diacritic.

The OCR programs that process the Arabic language were available in the market since more than 15 years ago, one of the old player in the market is Automatic Reader OCR from SAKHR who claimed that its accuracy can reach up to 99%. This is true that its accuracy can reach 99% if the fonts of the text is one of the new fonts used such as in MS WORD processor and this is true for all Arabic characters in Persian, Kurdish, Urdu and Bangali languages, but this rate will decrease to 50 to 60% when it comes to old Arabic printed texts because the letters could be overlapped for formatting or for esthetic reasons.

The Gold Edition of Atomatic Reader in Arabic (11.0) uses the Technology of Arabic Natural Language Processing (NLP) which recognizes the Arabic letters with accents and diacritics. The Gold Edition is characterized by the following features:

- Recognizes the diacritics in Arabic images

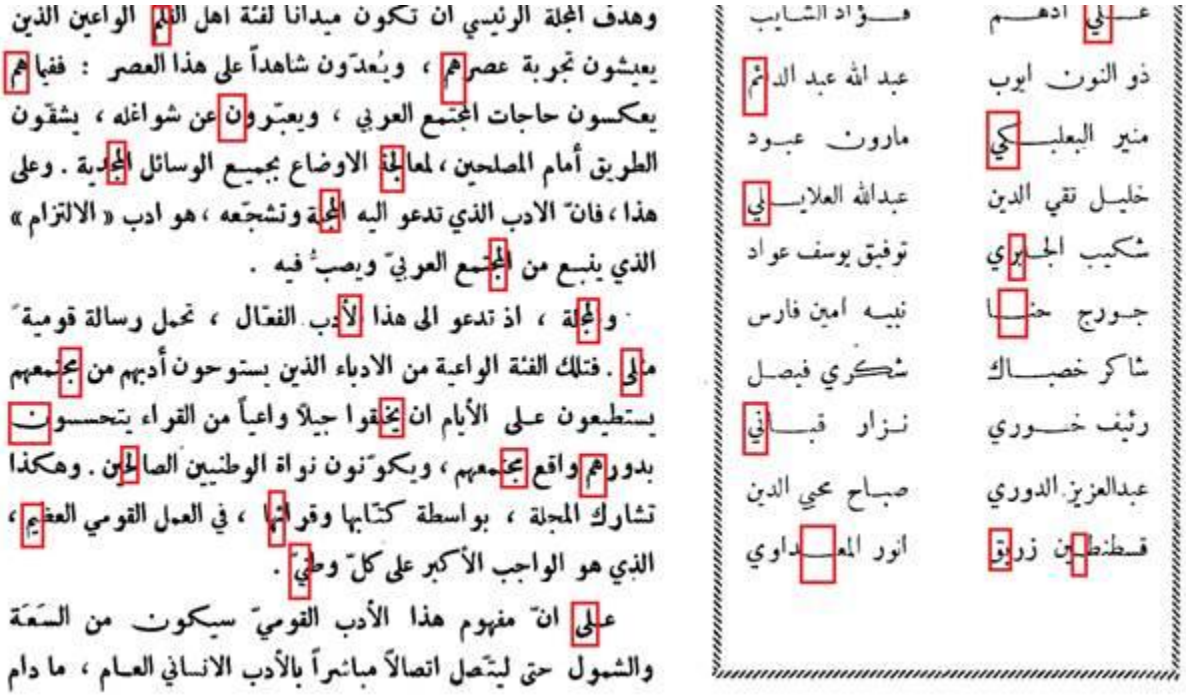
- Opens multiple documents at the same time
- Recognizes tables in scanned images
- Recognizes underlined words
- Recognizes broken and stickled characters with kashida
- Detects automatically style for fonts (Regular or Bold)
- Uses Arabic linguistic rules with recognition (Artificial Intelligence)
- Supports non-rectangular frames
- Supports color documents
- Groups recognition attributes into pre-defined types of source documents
- Includes bilingual spellchecker.
- Supports both automatic and manual framing modes.
- Enables zoning of text within frames in the pages. It detects the article of a newspaper by recognizing the blocks of texts within he white frames in the page.
- Keeps the initial formatting of the text.
- Enables morphological analysis of the form of the words and the structural analysis. It relies on the frequency of the words in the text.
- Deals with different image formats (.bmp, .tiff, .pcx, etc.)
- Saves the output text in different formats such as .txt, .rtf, and .html
- Supports PDF formats.

For the digitized Al-Ādāb Magazine Archive, several fonts were created by instructing the OCR SAKHR software to recognize and learn those fonts to cater for the different characters and printing that continuously changed since early years in 1950s till end of year 2012.

For each created font and applied to a new volume the accuracy rate was calculated at the character level which is the percentage of characters that were correctly recognized by the OCR engine compared to the total characters in a specific page. This facilitated tracking the learning curve for each font by SAKHR application where each time a new learning cycle was done the accuracy rate was re-calculated to check if it increased. This was repeatedly done until the learning process through the OCR led to a stable accuracy rate.

The accuracy varied from 80% to 99% in "Al-Ādāb" depending on the fonts, where old volumes had less font accuracy rate than the new volumes.

Of course the learning process and the accuracy rate were applied and calculated to a sample number of pages in each volume. The pages were selected thoroughly containing lot of text and representing well the different characters used in each volume. After reaching a stable accuracy rate, a batch process using SAKHR application was conducted to OCR all the volumes and to save the output of each image as pdf in a separate file.



وهدف المجلة الرئيسي ان تكون ميدانا لفئة اهل القلم الواعين الذين يعيشون تجربة عصرهم ، ويُعدّون شاهداً على هذا العصر : ففيهم يعكسون حاجات المجتمع العربي ، ويعبّرون عن شواغله ، يشقّون الطريق أمام المصلحين ، لمعالجة الاوضاع بجميع الوسائل المجدية . وعلى هذا ، فانّ الادب الذي تدعو اليه المجلة وتشجّعه ، هو ادب « الالتزام » الذي ينبع من المجتمع العربيّ ويصبُّ فيه .

والمجلة ، اذ تدعو الى هذا الأدب الفعّال ، تحمل رسالة قوميةً ملحّة . فتلك الفئة الواعية من الادباء الذين يستوحون أدبهم من مجتمعهم يستطيعون على الأيام ان يخلقوا جيلاً واعياً من القراء يتحسسون بدورهم واقع مجتمعهم ، ويكوّنون نواة الوطنيين الصالحين . وهكذا تشارك المجلة ، بواسطة كتّابها وقرائها ، في العمل القومي العظيم ، الذي هو الواجب الأكبر على كلّ وطنيّ .

على انّ مفهوم هذا الأدب القوميّ سيكون من السعة والشمول حتى ليتّصل اتصالاً مباشراً بالأدب الانساني العام ، ما دام

علي ادهم — فؤاد الشايب
ذو النون ايوب — عبد الله عبد الدائم
منير البعلبكي — مارون عبود
خليل تقي الدين — عبدالله العلايلي
شكيب الجابري — توفيق يوسف عواد
جورج حنا — نبيه امين فارس
شاكر خصباك — شكري فيصل
رئيف خوري — نزار قباني
عبدالعزيز الدوري — صباح محي الدين
قسطنطين زريق — انور المعداوي

Image 2.1 It shows overlapped characters, plus it shows Kashida, and how letters might vary in writing for example letter "ن".

4. Content Management System

AUB UL adopted the XTF (eXtensible Text Framework) as a content management system. It is an open source developed and maintained by California Digital Library to display its digital collections. Regarding Al-Ādāb Magazine Archives, the generated pdfs from the OCR engine in addition to the created metadata were uploaded to the XTF system. This of course required the implementation of several scripts to prepare the content and validate it against the prepared metadata. One of the main scripts was to merge the generated single page pdfs to create the articles as described in the metadata. This will facilitate later on searching for articles, viewing them and even downloading them online.

III. Challenges

1. Censorship and Publishing

Al-Ādāb Magazine was a liberal literary periodical in the Arab countries. It was open for the new creative ideas and doctrines that emerged in Europe after the 2nd World War. The editor-in-chief Dr. Souhail Idriss was open to publish for the intellectuals that were considered

supporting opposition in their own countries which created problems for the al- Al-Ādāb Magazine to be accessible in some Arab countries where the censor removed the opposition articles. The Magazine opted to publish two editions for the same issue, one is regular issue and the other is the censored issue. This solution created discrepancies in the content and in the numbering of pages in the same issue. It was also a challenge for the indexers who were obliged to create two different citations for the same article with a qualifier to differentiate between them and to add notes and cross-references to explain these discrepancies mainly at the level of the number of page continuity in order to ensure the access to the whole content of the issue.

2. Technical Challenges

A- Image capture & Enhancement

As mentioned previously, the existence of two different copies for the same issues led to a confusion during the scanning phase especially that the volumes in early years contained around 1,200 pages each. In addition, due to the missing of some covers for some issues in both copies, AUB UL had to look sometimes for a third copy.

During the scanning and as indicated in the methodology section all the content except the cover text were scanned in black and white, as this choice was to improve the OCR accuracy it had a bad result on the quality of the photos of the magazine. Accordingly, all the photos were scanned in gray scale and during the imaging process phase they were integrated into the digitized image.

Given that old volumes had bad quality paper compared to recent volumes/issues, this generated extra noise during the scanning phase which needed special treatment during the image processing phase to avoid extra dots that might lead to confusion during the OCR process. In this regard, some issues/pages contained very bad quality of the characters from the original document which impacted negatively the OCR accuracy.

Finally, in some volumes dot shading removal technique were used to clear the text for the OCR process and hence increase the accuracy rate.


صاحبها ومديرها المسؤول

الدكتور سهيل إدريس

سكرتيرة التحرير

عايدة مطرجي إدريس

الإدارة

شارع سانية الجنزير

بناية بيهم - تلفون ٨٦١٦٣٣ - ٨٠٣٧٧٨

تلكس ADAB 23402LE ص . ب ٤١٢٣ - بيروت، لبنان

Image 3.1: Example of Dot Shading area

B- Preparing the descriptive metadata

At the early stages of the digitization process it was agreed to use the Library information system to index the data as because this will allow the generating of metadata in different formats including the MARC/MARCXML format that will help the integration of metadata in different system later on.

Despite the problem of censorship discussed at the beginning of the challenges section, the librarian who was preparing the metadata has faced technical problems regarding the publishing frequency of the magazine in the bibliographic citations which varied from monthly to bi-annual to quarterly at the beginning of the year 2012. In addition, the number of pages in the annual volume varies from 400 to 1000 pages for the issues in the whole year This was due to the events that happened in Lebanon and the Middle East and prevented writers to send their articles to the publisher. The main concern in this change of frequency that it affects the number of digit for the issue number. The following examples can highlight the citation problems which created problems in the structure of the database definition.

v.01 no.01(1953) was in the 1st year

v.24 no.01-03 (1976) was in the year 1976

v.60 no. Spring (2012) was in the year 2012

In addition, the rubric of the categories of the magazine changed over time causing some complexity at the cataloguing level such as the concept of “File” that contained several related articles. Therefore, it was after cataloguing several issues from the past and from the new era the global information required in descriptive metadata was obvious and structured.

Another issue was the cataloguer/indexer needs to identify continuation of the articles in order to link the corresponding pages accordingly. This created extra work and required high attention as some articles span on consecutive pages while others start on a page and continues at the end of the issue.

Finally, in addition to the need of descriptive metadata to search and view information about items in this digital collections, the metadata also improve the search results and compensate for the low accuracy of OCR especially in old volumes.

C- OCR Process

In addition to already known problems and complexity in OCRing Arabic Text where some of these issues were raised in the methodology section such as overlapping, there are several others issues related to Arabic characters such as:

- Some characters have the same form and are only distinguished by the position of various dots relative to the main character block. Given that dots are considered as noise and OCR tends to remove them.
- Space between two connecting Arabic characters can vary in size and shape. e.g.

سـعيد أو سعـيد أو سعيـد أو سعيد

Vowels: They are different from diacritical marks examples fatḥa, dhammah, kasrah (short vowels) e.g.

قبِل أو قُبل أو قِبل

However, another problem appeared when OCRing Al-Ādāb, it is related to the inconsistency of fonts size and calligraphy. This is not to refer to Title of articles, author names or other but this occurred mainly in old volumes where several “Fonts” were used in the same volume (year), in the same issue and even in the same page as show in the image.

أمين مشرق : الأديب الشاعر المجهول

بقلم حاتم طه الراوي

الشاعر العبقري المجهول الذي صان ماءَ جبينه ، وسقى دوحة شاعريته بدمع عينه وعبرات فؤاده ، هو أسمى وأرفع قدراً من ألف أمير للشعر يغمس يراعه في قلوب غيره من الشعراء ثم يساوم ذوي الجاه والسلطان ليغمروه بالعطايا والصدقات .

والأديب العبقري المجهول ، الذي يغترف أدبه من ينبوع الحياة ، من شعبه الحائر ومن شعوب الأرض الحائرة المتألمة المستطلعة الى مستقبل أفضل ، هو أسمى وأرفع قدراً من أديب ينثر ما ينظمه الاخرون ، ويزخرفه ، ثم يتملّق هذا ويجثو امام ذاك ، طمعاً في الحطام ، حطام الدنيا الزائل الذي كان في الماضي البعيد والقريب لطخة عار في جبين الأدب العربي . إنه يتسلّق سلم الشعوذات قاصداً الشهرة ، والشهرة قد تكون في بعض الاحيان كالموس من استرضاها كان دونها قدراً كما يقول ميخائيل نعيمة ...

فالشهرة ليست ضرورية للأديب ، بل قد تكون وبالاً عليه عندما تحيطه بهالة من التقديس ، ودونك ما فعلته الشهرة العريضة بكبار ادبائنا المعاصرين في إنتاجهم الرخيص ..

والمرحوم أمين مشرق كان شاعراً مجهولاً وكاتباً غير معروف الا لدى هواة الأدب المهجري ، ولم ينصفه أحد من أدباء العربية سوى الدكتور محمد مندور في كتابه النفيس « في الميزان الجديد » .

وقد طلبتُ من صديقه ميخائيل نعيمه عام ١٩٥٠ أن يزودني ببعض المعلومات عن حياة هذا الشاعر الفريد ، وإذا بمعلومات نعيمة ضئيلة أيضاً في هذا الموضوع فلقد كتب لي يقول : « المرحوم امين مشرق كان صديقي ولكني لا اعرف الكثير عن حياته ، الا انه هجر الى الولايات المتحدة اولاً ومنها الى الأكوادور . ثم عاد الى لبنان عام ١٩٣٢ حيث كانت الفرصة الأولى لتعارفنا . وتزوج في العام ذاته وعاد الى الأكوادور . وبعد عام او عامين قضى في حادث سيارة . أما آثاره الشعرية — على قلتها — فأكثرها من طراز ممتاز ... »

★

لم يخلّف المرحوم امين مشرق كتاباً نثرياً ولكن مقالاته التي نشرها في صحف المهجر « كالسائح » وغيرها تغني عن كتب كثيرة . وقد كانت مقالاته صرخات داوية في أذن الرجعية العمياء . فقد كان الأمين مجدّداً في الوقت الذي كان فيه التجديد كفراً لا يُغتفر للأديب ، وفي الفترة التي كانت فيها الافكار التقدمية مثاراً للسخرية والاشمئزاز ومطّ الشفاه !..

ولم يكن امين مشرق منجرفاً مع تيار التجديد ، لأنه لم يكن من الذين ينجرفون مع تيار النزعات الفكرية من غير تفكير ، ذلك انه لم يكن إمّعة شأن بعض المهرجين الذين لا يتورعون عن وضع سقط المتاع في آنية التجديد المزيّف . لأن النزعة التجددية كانت ممتزجة مع دمه ، بل كانت عواصف تزأر في أعماقه فيطلقها يراعه ...

وحسبك أن تقرأ له بحثه الطويل « الداء العياء » المنشور في كتاب « ما وراء البحار — أو النبوغ العربي في العالم الجديد » لتوفيق الرافعي . فقد أوضح فيه الفقيد أدق المقاييس التي يجب أن يقاس بها الأدب الحي . وقد استهل بحثه بتأييده لميخائيل نعيمة في انتقاده اللاذع لدرّة من درر شاعر الأمراء المرحوم احمد شوقي . ثم طفق يندد بالمبالغة المقيتة التي طغت على أدبنا القديم وسرت عدواها الى أدبنا الحاضر قائلاً :

« لو سمعنا أحد الشعراء يرثي اسكافاً مات بين النعال والاحذية قائلاً ان الفضل مات بموته والعلم هد ركنه والأدب امسى يتيما ، ويتعجب كيف ان النجوم لم تنطق حداداً والدهر لم يقف جائراً . أو لو قرأنا شعراً لآخر يمدح فيه انور باشا وحصانه الأدهم بقوله: ان صهيله «في قلب اوربا له ترديد» . أو لو سمعنا عاشقاً ينشد :

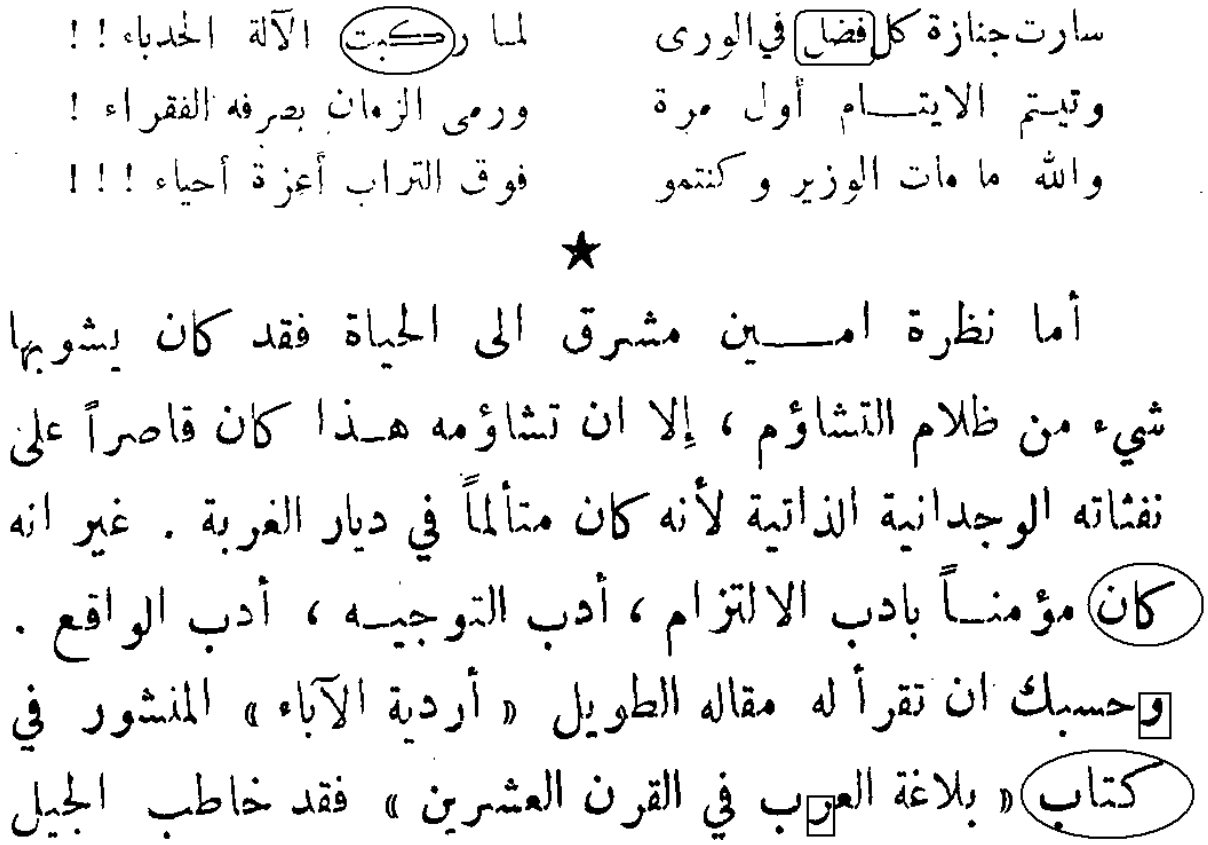
أمر بالحجر القاسي فألثمه ... لأن قلبك قاس يشبه الحجرا

وسألنا الثلاثة لماذا كان هذا الغلو ، لضحكوا منا ولا شك مشفقين لجهلنا ثم اخرج اولهم من تحت ابطه كتاب علم المعاني والبيان. وأظهر الثاني ديوان المتنبي أو ابن الفارض. وفتح الثالث كتاب نهج البلاغة، وقدموها الينا وقد لاحت ابتسامة الانتصار على ثغورهم ولسان حالهم يقول : تعلموا هنا قواعد البلاغة وحدود البيان ثم لا تحتاجون الى سؤال .. »

ثم يثور امين مشرق على هؤلاء ثورة مشفق ساخر قائلاً :

« هؤلاء القوم ويا للاسف معذورون بعض العذر . كيف لا وكل ما تعلموه منذ ان اصبحوا يتهاجون بالكلمات يبتدىء بـ « حدثنا سهيل بن عباد » وينتهي بشرح المعلقات السبع ؟ أيلام التلميذ على حفظه مسائله ؟ أو ليس طبيعياً ان تنمو النبتة معوجة اذا ربطناها بحائط معوج ؟ وهل الذنب ذنب الارض

Image 3.2: Different Font layout and size on a same page from Al-Ādāb issue


سارت جنازة كل فضل في الورى لما ركبت الآلة الخدباء ! !
وتيتم الايتـــام أول مرة ورمى الزمان بصرفه الفقراء !
والله ما مات الوزير وكنتمو فوق التراب أعزة أحياء ! ! !

★

أما نظرة امـــين مشرق الى الحياة فقد كان يشوبها شيء من ظلام التشاؤم ، إلا ان تشاؤمه هـذا كان قاصراً على نفثاته الوجدانية الذاتية لأنه كان متألماً في ديار الغربة . غير انه كان مؤمنـاً بادب الالتزام ، أدب التوجيـه ، أدب الواقع . وحسبك ان تقرأ له مقاله الطويل « أردية الآباء » المنشور في كتاب « بلاغة العرب في القرن العشرين » فقد خاطب الجيل

Image 3.3: The letter "ك" written in three different ways in the same page even in the same paragraph.

Accordingly calculating the accurate was not straightforward, sometimes three different fonts were used for the same volume and the accuracy rate was calculated for each font separately. When the fonts were mature we had to create a library which is a feature in SAKHR application that allows to detect multiple fonts.

D- Open Source Content Management System

Despite the benefit of using the XTF open source system, the main challenge was to prepare the content to be uploaded into the system. Several scripts were created to validate the data and the content prior to uploading the data in order to minimize error that could be the result of manual processing thousands of digital files and records. For example, some scripts made sure that each generated article is linked to a record in the metadata or all records in the metadata are linked to a pdf file. After checking the consistency of the files that are uploaded to the system where it was customized for search to specific files in addition to full text search.

To solve the problems of articles that start on one page and finish on other the pdfs were merged into a single downloadable pdf.

IV- Summary

Digitizing Al-Ādāb Magazine Archives generated around 50,000 digital files from 60 volumes. All digital files were saved for digital preservation based on AUB UL policies and procedures. The indexing/cataloguing phase led to the compilation of around 15,000 articles that were saved in pdf format with a full descriptive metadata record for each article (pdf). The detailed descriptive metadata compensated for the low OCR accuracy especially for old volumes. In general, around 10 fonts were created to cover the different characters in different Al-Ādāb volumes/issues published throughout 60 years with an average of OCR accuracy ranging from 80% till 99%.

The adopted methodology for digitizing and indexing the Al-Ādāb magazine was based on recommended standards and best practices adopted by leading institution in this area and the proposed methodology is flexible to accommodate new type of items/collections that need to be digitized and published online to be searchable, viewable and downloadable.

Finally, similar projects might face same problem or new problems specific to the digitized collections. The importance is to examine, assess and evaluate the materials of the collection before engaging in any digitization process to anticipate any issues that will hinder or imply additional time in the digitization workflow to better balance between quality and efficiency.

References

- Calhoun, K. (2014). Exploring Digital Libraries Foundations, practice, prospects.
- Smith, A.(2001). Strategies for Building Digitized Collections.
- http://al-adab.com/article/الآداب-في-عهدها-الثالث-ما-يَمكث-في-الأرض-ويحلّق-نحو-السماء
- http://www.al-adab.com/archive (أرشيف الآداب ١٩٥٣-٢٠١٢)
- Federal Agencies Digitization Guideline Imitative. Web Technical Guidelines for Digitizing Cultural Heritage Materials (2010) http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf
- Automatic Reader 11.0, GoldEdition, Arabic OCR. Retrieved from http://aramedia.com/ocr80gold.htm (26 January 2016)
- The XTF (eXtensible Text Framework) http://xtf.cdlib.org/about (26 February 2016)
- The XTF (eXtensible Text Framework) http://xtf.cdlib.org/ (26 February 2016)