A tool for semiautomatic cataloguing of an islamic digital library: a use case from the Digital Maktaba project (short paper)

Riccardo Martoglia[1],*, Luca Sala[2], Matteo Vanzini[2] and Riccardo Vigliermo[1]

[1]*University of Modena and Reggio Emilia, 41125 Modena, Italy*
[2]*mim.fscire, 40125 Bologna, Italy*

Abstract

Digital Maktaba (DM) is an interdisciplinary project to create a digital library of texts in non-Latin alphabets (Arabic, Persian, Azerbaijani). The dataset is made available by the digital library heritage of the "La Pira" library in the history and doctrines of Islam based in Palermo, which is the hub of the Foundation for Religious Sciences (FSCIRE, Bologna). Establishing protocols for the creation, maintenance and cataloguing of historical content in non-Latin alphabets is the long-term goal of DM. The first step of this project was to create an innovative workflow for automatic extraction of information and metadata from title pages of Arabic script texts. The Optical Character Recognition (OCR) tool uses various recognition systems, text processing techniques and corpora in order to provide accurate extraction and metadata of document content. In this paper we address the ongoing development of this novel tool and, for the first time, we present a demo of the current version that we have designed for the extraction and cataloguing process by showing a use case on an Arabic book frontispiece. In particular, we delve into the details of the tool workflow for automatically converting and uploading PDFs from the digital library, for the automatic extraction of cataloguing metadata and the semiautomatic (at the current stage) process of cataloguing. We also shortly discuss future prospects and the many additional features that we are planning to develop.

Keywords

Cultural heritage, Digital Library, Islamic sciences, Arabic script OCR, Information extraction, Output alignment, Page layout analysis, Semiautomatic cataloguing, Software tool usage demo.

1. Introduction

Creating methods and procedures for preserving and making cultural heritages accessible in multilingual contexts is one of the new and pressing demands placed on content managers by the necessity to manage multimedia content today. This is the difficult situation that the Digital Maktaba Project (DM) is attempting to address. DM was possible through a gathering of expertise from the start-up mim.fscire, the University of Modena and Reggio Emilia (UniMoRe) and the Fondazione per le Scienze Religiose (FSCIRE), leader institution of the RESILIENCE European research infrastructure on Religious Studies (ESFRI Roadmap, 2021). The fundamental objective of DM is to provide a cutting-edge tool and methods for the management of multilingual literary patrimony, paying special attention to literature written in scripts other than Latin. The design

Qurator 2022: 3rd Conference on Digital Curation Technologies, September 19-23, 2022, Berlin, Germany

riccardo.martoglia@unimore.it (R. Martoglia); salaluca.info@gmail.com (L. Sala); vanzinimatteo.info@gmail.com (M. Vanzini); riccardo.amerigo.vigliermo@unimore.it (R. Vigliermo)
0000-0003-4643-6128 (R. Martoglia)

of a novel process and application is currently focused on the automatic knowledge extraction and categorization of texts written in non-Latin scripts, particularly those written in the Arabic, Persian and Azerbaijani languages. The vast collection of digital publications made internally accessible by the "Giorgio La Pira" library in Palermo, a hub of the FSCIRE organization and devoted to the history and teachings of Islam, serves as the project test-case.

This paper proposes a demo of the DM software we are currently developing. In particular and for the first time, we describe a detailed use case scenario and show how the current version of the software, which builds on the innovative foundations that have been described in previous works [1, 2], can already significantly help the cataloguing of non-latin works beyond the tools that are currently available in state of the art.

In Section 2, we will discuss some related works. In Section 3, we take a look at the current prototype by proposing an overview of the tool features, from the digitization[1] of documents to the cataloguing process phase. Section 4 presents a use case on a PDF document in Arabic language drawn from the library. Finally, Section 5 concludes the paper by providing a look to the future perspectives and features we are planning for the tool, including the use of machine learning (ML) techniques in order to make its usability and effectiveness grow more and more with its usage.

2. Related Works

Even though the Natural Language Processing (NLP), Information Retrieval (IR) and OCR fields on Arabic scripts have made huge strides in the last decades, there have been not many projects aimed at exploiting them for curating new and innovative digital libraries. Compared to the other automation issues (automatic cataloguing with ML techniques, automatic extraction of metadata, etc.), as far as we know, there are very few significant proposals within the languages considered. Recent similar projects concern the digitization and the building of Arabic and Persian text corpora. One is represented by the Open Islamicate Text Initiative (OpenITI)[3] which is a multi-institutional effort to construct the first machine-actionable scholarly corpus of premodern Islamicate texts. OpenITI contains almost exclusively Arabic texts, put together into a corpus within the OpenArabic project. The arabic texts are collected from open-access online libraries such as Shamela [4] and Shiaonline library [5]. From OpenITI, two other interesting projects have been developed: KITAB [6] at the Agha Khan University and the Persian Digital library (PDL) at the Roshan Institute for Persian Studies of the University of Maryland [7]. The first one provides a toolbox and a forum for discussions about Arabic texts and its main goal is to research relationships between Arabic texts and discover the inter-textual system laying underneath the Arabic rich textual tradition. The latter is focused primarily on the construction of a scholarly-verified corpus and secondarily on the development of an OCR system for handwritten Persian texts. PDL has already created an open-access corpus of Persian poems collected from the Ganjoor site [8] and then integrated with a lemmatizer [9] and a digital version of the Steingass persian dictionary [10]. From a Character recognition standpoint, the

[1] With digitization we refer to all the operations involved in making a document retrievable and searchable, from the basic scanning to the post capture processing. Therefore, a scanned document is not necessarily a *digitized* document.

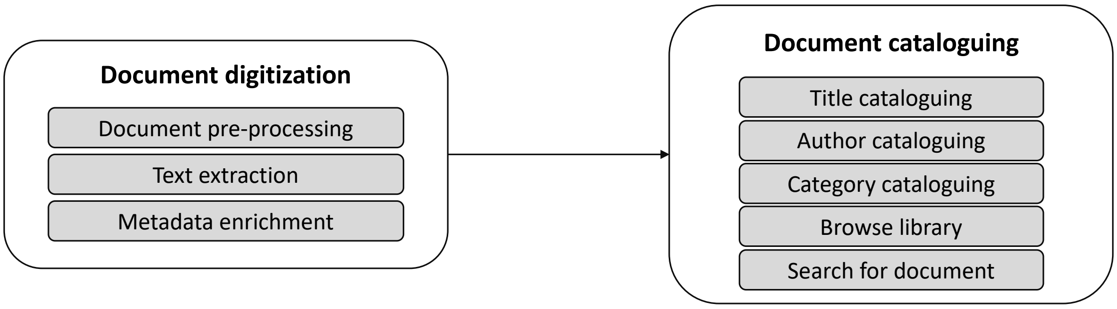


Figure 1: DM tool's main workflow, detailing the two macro-activities performed in the software

OpenITI project sees the employment of Kraken OCR useful both for handwriting and printed text recognition [11, 12]. All the projects we have discussed focus only on a limited subset of the languages taken into account by DM and aim to completely digitize a (relatively) small library of books, frequently requiring a large amount of manual work. Moreover, among the unique features that are already implemented in the tool (see also next section), DM provides support for extracting a very rich array of metadata (syntactic, linguistic and cataloguing) rather than text content alone, fully supporting non-latin alphabet metadata and without manual work required from the user.

3. DM software overview

The DM tool workflow can be divided into two main logical macro-activities (see also Figure 1), corresponding to the two main components of the software:

- The first, devoted to the **digitization of documents** and explained in Section 3.1, includes all the operations that are carried out to prepare and process the documents to be catalogued by the user: document pre-processing,text extraction performed by means of a novel approach exploiting several OCR engines and metadata extraction / enrichment;
- **Document cataloguing**, detailed in Section 3.2, is the second macro-activity supported by the tool through its web interface and intelligent features. To make this cataloguing tool as accessible as possible, a web application was created using the Python programming language and the Flask framework, with the goal of providing a simple and intuitive interface, accessible anywhere and from any device to the users who will catalogue the documents and whoever has the need to search the library database afterwards. All data from the catalogued documents are saved and stored in PostgreSQL database[2].

3.1. Digitization of documents

Since the collected material comes from different sources and it is composed of very heterogenous files, a common document pre-processing phase has been created to manage all types of

[2] https://www.postgresql.org/ (Accessed on 25 July 2022)

input and help filtering out those that are not interesting for the purposes of the cataloguing. Documents are split between text-searchable and non-text-searchable ones; OCR techniques must be employed to extract image content from non-text-searchable documents since they lack editable text. As to the actual text extraction phase, simply using publicly available OCR tools was not an option in DM: while there are certainly good proposals for Latin script languages, the same thing cannot be said for Arabic script languages. For DM, we tested several open source OCR systems (Tesseract[3], EasyOCR[4] and GoogleDocs[5]), aiming to combine and extend them to reach the best possible output by exploiting their specific advantages. Test results brought us to select EasyOCR and GoogleDocs as they proved to be more suitable for our task. While certainly a good starting point, these systems, when employed alone, had a number of drawbacks. For example, on one hand EasyOCR can extract small amounts of text with medium quality and is one of the few tools to provide some kind of metadata (even if limited to the location of the text in the original image); however, it requires manual language specification before processing. Moreover, text region metadata proved to be quite inaccurate, with text regions usually fragmented in too many boxes and not correctly arranged according to Arabic's right-to-left ordering. On the other hand, GoogleDocs offers automatic language detection and an improved output quality, however its output doesn't contain any kind of metadata.

In order to go beyond the above limitations, in DM we implemented a processing pipeline that combines these libraries in a novel and comprehensive manner to obtain a richer and a higher-quality output. The DM pipeline also includes a novel text region grouping, merging and renumbering phase that follows the OCR processing and precedes the cataloguing phase to enhance metadata usability.

Together with the *syntactic* metadata coming from text extraction, additional *linguistic* metadata is extracted in the metadata enrichment/extraction phase. Searching for effective linguistic resources covering the project languages in the open source environment was another challenging task. There aren't any open source linguistic resources for Arabic, Persian or Azerbaijani that have coverage levels even close to that of the English language as of now. We made the decision to base our tool on multiple resources in order to partially address this problem: the Open Multilingual WordNet thesauri including Arabic and Persian WordNet[6], Tashaphyne[7] and Arramooz[8]. DM also automatically generates and stores *cataloguing* metadata, through a novel approach for the automatic title and author identification in a frontispiece. More details of the techniques can be found in [2], while an exemplification of their functioning on an actual document cataloguing use case will be described in the use case demo (Section 4).

3.2. Cataloguing process and interface

In the cataloguing process, the cataloguing UI and features are exploited in order to assist the librarian (see also Section 4 for all the details on how the interface is structured and employed).

[3] Tesseract. Available online: https://github.com/tesseract-ocr/tesseract (Accessed on 20 July 2022)

[4] EasyOCR. Available online: https://github.com/JaidedAI/EasyOCR (Accessed on 20 July 2022)

[5] GoogleDocs. Available online: https://docs.google.com (Accessed on 20 July 2022)

[6] Open Multilingual WordNet thesauri. Available online: http://compling.hss.ntu.edu.sg/omw/ (Accessed on 21 July 2022)

[7] https://pypi.org/project/Tashaphyne (Accessed on 21 July 2022)

[8] https://github.com/linuxscout/arramooz (Accessed on 21 July 2022)

Once document digitization is completed, the interface shows the first page of the elaborated document (cover or frontispiece) with the extracted text regions (properly merged and reordered). Manually adding the title, authors and other information for cataloguing a new document is one of the most crucial but burdensome tasks for a librarian. In this phase, DM makes use of the extracted text and metadata to automatically suggest to the librarian the text sections most likely to contain particular fields (at this time, title and author). Then, other selections can be performed, including topic or category which the text belongs (e.g. Qur'an, Tafsir, Islamic law, Islamic philosophy etc.). Finally, the document is catalogued and all its information are stored in the library database. At this point the document is retrievable by inserting one of the cataloguing metadata whether is the author, the title, the category or a combination of these.

4. Cataloguing use case and demonstration

In this section, the complete cataloguing process of a sample document is shown. Starting from an uploaded folder containing the document itself, the system elaborates the document by preparing it to the cataloguing phase. Next, the document is ready to be catalogued through the title, author and category pages. At the end of this process, the document will be searchable and retrievable by using the search page. In particular, our use case focuses on a Monography on History of Islamic civilization and specifically on the period of the Caliph Yazid bin Mu'awiyya. From a graphic-linguistic point of view an arabic character is employed in all the frontispiece entities (naskh); the presence of an image on the frontispiece will test the tool ability in recognizing and selecting text portions from composite pages. Figure 2 covers the different actions (and respective UI pages, each denoted with a capital letter) performed in the use case and described in detail in the following.

A. Uploading a folder of documents. The first step, as shown in Figure 2 (A), requires the user to select the folder containing the PDF documents (s)he wants to upload on the web application. Once selected, using the "upload" button, the system will transfer them to the server and start processing. Currently, processing takes place via a priority-based queue. The documents go through multiple steps: preparation of an *ad-hoc* folder on the server, conversion of PDF pages to PNGs, processing of the title page through OCR engines and finally processing the metadata as explained in Section 3.1.

B. Main page. The processing of documents uploaded by the user takes place in the background. This allows the tool to process them without interrupting the main functioning of the web application. In this way, every time a document has finished the pre-processing phase and is therefore ready for cataloguing, it will be immediately shown in the home page. Meanwhile, the other documents will continue to be prepared in the background. As shown in Figure 2 (B), on the main page of the DM web application it is possible to understand if the system is still processing documents thanks to the presence of a loading icon.

The cataloguing of documents consists of three steps: title insertion, author insertion and category selection. The system requires users to go through the three steps in order but also allows them to resume from an intermediate step in case of breakdowns or interruptions. Figure 2 (B) shows three corresponding sub-sections in the main page, containing information on documents ready for title, author and category cataloguing. Each document will appear in the first sub-section immediately after the pre-processing phase, while in the following ones it will be displayed only if the cataloguing is interrupted.

C. Title cataloguing. Figure 2 (C) shows that in this phase the screen is divided into two columns: in the left one, the images of the pages that compose the document (in this case, the one from our use case) are displayed, while the right column shows the information extracted from the images. By default the displayed page is the first one, but it is possible to change it using the arrows positioned on the right and left of the image. Text information is divided into boxes that are displayed on the image itself (left column), while in the right column there is the title entry box and the information extracted using OCR, presented in an ordered list in an information pane. By doing this the text and its location within the page are visually corresponding.

In the case of our use case, we can see that text is extracted correctly. All the information is extracted by means of the processing workflow outlined in the previous section. In particular, text regions are properly merged and reordered allowing to drastically reduce wrongly merged or sorted boxes by order of magnitudes w.r.t. state of the art (e.g., from our tests on several documents, from 30% to less than 2% of wrongly sorted boxes).

As soon as the title entry page is shown, the system automatically proposes a plausible box containing the title, exploiting a number of heuristics based both on syntactic (e.g., text size and position) and semantic/linguistic information. For our document the title is extracted and hinted correctly. In general, the accuracy of this feature is currently around 70% and will certainly be improved in the future, however note that none of the state of the art systems aims at automating this task, thus requiring completely manual insertions. In case the box is wrongly selected box or its content is incorrect, the user can choose to make changes to the text or to change the selection. Text editing is performed by simply correcting the characters placed in the title input box. The selection can be made either by clicking on a box positioned on the image or on a box in the information pane.

If the title is somehow fragmented into several boxes, it is possible to select them in order to reassemble it. Once done, the text contained in the selected boxes will be aggregated (in the selection order) and the user will still have the possibility to correct it in the input box. In this regard, a notable thing about DM is that it is designed to store not only the final user-approved information, but also the originally identified text and how it (possibly) had to be modified by the user: this will enable future ML techniques that will be able to enhance the system effectiveness through use.

A further visual aid for the user is given by the implementation of multiple linguistic resources. By using these, the system searches the extracted text for a word-by-word correspondence to check the presence of a word in the vocabulary (by means of green color). By clicking on the paperclip the user can view the detailed linguistic information retrieved by the system. The overall linguistic coverage achieved by using more than one resource is in the order of 75%, much higher than what could be obtained by using a single resource (e.g., 10% for WordNet).

D. Author cataloguing. In the author cataloguing step (Figure 2 (D)), the web application interface is similar to the title cataloguing one, with some slight differences. The system suggests a box in which it is confident the author's name and surname are present; again, for our use case the author is extracted and hinted correctly. As for the text inserted in the input box, in this case it is formatted in the form "surname, name". When there are multiple first names and last names, which is common in the Arabic language, it is sufficient to combine them before and / or after the comma (e.g. "surname, name_one name_two" or "surname_one surname_two, name").

Another difference from the previous step is the information related to the text shown. In this case, word-by-word matches are searched within an *ad-hoc* dataset constructed by combining external source information (i.e., lists of Arabic names and surnames), and matching names are highlighted in green.

E. Category cataloguing. After entering the title and author of the document, the category is selected. As shown in Figure 2 (E), in the current version categories must be manually chosen by the user from a 3-level hierarchy of more than 560 entries, according to the topographic design of the La Pira library. In the future, we intend to offer "smart" AI-based ways to help the process by offering pertinent cataloguing suggestions. The user can enter the first-level category through the drop down menu and then select, if applicable, the second-level category and third-level category in the same way. In the case of our document, the PDF is categorized in the History of Isalmic societies section (Unitl 1258).

F. Browse library page. Once the document has been catalogued, it can be viewed within the "Library" page, Figure 2 (F), which contains all the documents ordered from the most to the least recently inserted. Through this page it is possible to view, modify or delete documents.

G. Search for catalogued documents. The "Search" page, Figure 2 (G), offers the possibility to retrieve each catalogued document through a simple search by specifying a single parameter, or through an advanced search by specifying more than one. Several kinds of indexes enable DM querying, in particular Generalized Inverted Indexes supporting title and author full-text search and B+trees for category lookup. The results of the search is displayed to the user as illustrated in Figure 2 (H).

5. Conclusions and Future Work

In this paper we presented a demo / use case on the first version of the DM tool for the automatic extraction of knowledge from documents written in multiple non-Latin languages.

The tool is still far from finished, we conclude the paper by briefly discussing some of the research directions, features and advantages we foresee for its future versions:

- **data interchange and long-term preservation aspects** will be investigated in order to allow data interchange with catalogue data from other libraries and make the managed data readable and usable also on a long-term basis;
- **intelligent and AI-based techniques** will be designed and developed in order to bring new levels of assistance to the cataloguing process. Automatic suggestions will be improved through the study of user feedback and previously entered data. The automation of tasks such as the automatic recognition of the type of publication and the classification of the document itself according to the project of the Giorgio La Pira library will be implemented through the study of supervised ML models. The development of incremental ML algorithms will ensure that the tool can "learn," becoming more automatic and effective with use. Both traditional and deep learning methods will be studied and they will be

implemented on parallel architectures for faster execution. A special focus will be placed on interpretable ML algorithms in order to advance beyond the black box aspect of ML proposals and explain them, while also being in line with the current ML trends popular on other domains. Moreover, from a library cataloguing / cultural heritage context, this would be something that has seldom been done in current projects.

All the techniques mentioned above will be combined to build an increasingly automated tool, enabling a simpler and faster cataloguing procedure and, eventually, better cultural heritage preservation and dissemination.

References

[1] S. Bergamaschi, R. Martoglia, F. Ruozzi, R. A. Vigliermo, S. D. Nardis, L. Sala, M. Vanzini, Preserving and conserving culture: first steps towards a knowledge extractor and cataloguer for multilingual and multi-alphabetic heritages, in: O. Gaggi, P. Manzoni, C. E. Palazzi (Eds.), GoodIT '21: Conference on Information Technology for Social Good, Roma, Italy, September 9-11, 2021, ACM, 2021, pp. 301–304. URL: https://doi.org/10.1145/3462203.3475927. doi:10.1145/3462203.3475927.

[2] S. Bergamaschi, S. De Nardis, R. Martoglia, F. Ruozzi, L. Sala, M. Vanzini, R. A. Vigliermo, Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The digitalmaktaba approach, Sensors 22 (2022). URL: https://www.mdpi.com/1424-8220/22/11/3995. doi:10.3390/s22113995.

[3] M. T. Miller, M. G. Romanov, S. B. Savant, Digitizing the textual heritage of the premodern islamicate world: Principles and plans, International Journal of Middle East Studies 50 (2018) 103–109. doi:10.1017/S0020743817000964.

[4] al-Maktabah al Shamela, Shamela library, Last accessed: July 18, 2022. URL: https://shamela.ws/.

[5] Shiaonline, Shiaonline library, Last accessed: July 18, 2022. URL: http://shiaonlinelibrary.com.

[6] A. K. U. International, Kitab project, Last accessed: July 15, 2022. URL: https://kitab-project.org/about/.

[7] U. o. M. Roshan Institute for Persian Studies, Persian digital library, Last accessed: July 18, 2022. URL: https://persdigumd.github.io/PDL/.

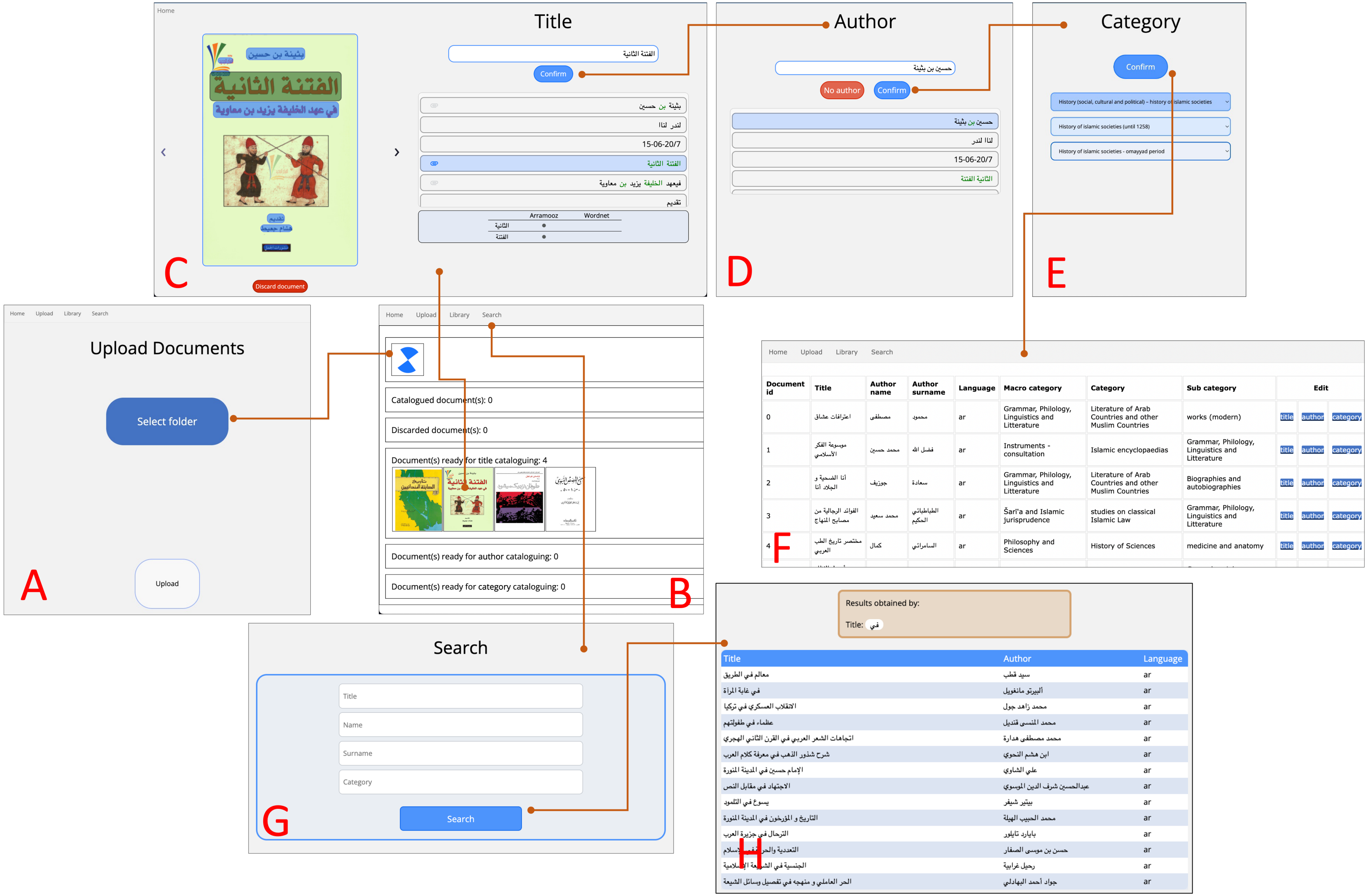
[8] Ganjoor.net, Ganjoor, Last accessed: July 18, 2022. URL: https://ganjoor.net/.

[9] Roshan-ai.ir, Hazm, baray-e pardazesh-e zaban-e farsi, Last accessed: July 13, 2022. URL: https://www.roshan-ai.ir/hazm/.

[10] F. J. Steingass, A Comprehensive Persian-English dictionary, including the Arabic words and phrases to be met with in Persian literature, Routledge & K.Paul, London, 1892. URL: https://dsal.uchicago.edu/dictionaries/steingass/.

[11] M. Romanov, M. Miller, S. Savant, B. Kiessling, Important new developments in arabographic optical character recognition (ocr), Al-'Usur al-Wusta 25 (2017). doi:10.7916/alusur.v25i1.6996.

[12] B. Kiessling, Kraken - a Universal Text Recognizer for the Humanities (2019). URL: https://doi.org/10.34894/Z9G2EX. doi:10.34894/Z9G2EX.



Document id	Title	Author name	Author surname	Language	Macro category	Category	Sub category	Edit		
0	اعترافات عشاق	مصطفى	محمود	ar	Grammar, Philology, Linguistics and Litterature	Literature of Arab Countries and other Muslim Countries	works (modern)	title	author	category
1	موسوعة الفكر الأسلامي	محمد حسين	فضل الله	ar	Instruments - consultation	Islamic encyclopaedias	Grammar, Philology, Linguistics and Litterature	title	author	category
2	أنا الضحية و الجلاد أنا	جوزيف	سعادة	ar	Grammar, Philology, Linguistics and Litterature	Literature of Arab Countries and other Muslim Countries	Biographies and autobiographies	title	author	category
3	الفوائد الرجالية من مصابح المنهاج	محمد سعيد	الطباطبائي الحكيم	ar	Šarī'a and Islamic jurisprudence	studies on classical Islamic Law	Grammar, Philology, Linguistics and Litterature	title	author	category
4	مختصر تاريخ الطب العربي	كمال	السامرائي	ar	Philosophy and Sciences	History of Sciences	medicine and anatomy	title	author	category

Title	Author	Language
معالم في الطريق	سيد قطب	ar
في غابة المرآة	أليبرتو مانغويل	ar
الانقلاب العسكري في تركيا	محمد زاهد جول	ar
عظماء في طفولتهم	محمد المنسى قنديل	ar
اتجاهات الشعر العربي في القرن الثاني الهجري	محمد مصطفى هدارة	ar
شرح شذور الذهب في معرفة كلام العرب	ابن هشم النحوي	ar
الإمام حسين في المدينة المنورة	علي الشاوي	ar
الاجتهاد في مقابل النص	عبدالحسين شرف الدين الموسوي	ar
يسوع في التلمود	بيتير شيفر	ar
التاريخ و المؤرخون في المدينة المنورة	محمد الحبيب الهيلة	ar
الترحال في جزيرة العرب	بايارد تايلور	ar
التعددية والحرية في الإسلام	حسن بن موسى الصفار	ar
الجنسية في الشريعة الإسلامية	رحيل غرايبة	ar
الحر العاملي و منهجه في تفصيل وسائل الشيعة	جواد أحمد البهادلي	ar

Figure 2: DM cataloguing use case: the letters correspond to the different UI pages and actions described in the text