Umanistica Digitale – ISSN: 2532-8816 – n. 22, 2026: Special Issue
R.A. Vigliermo, G. Sullutrone, S. Bergamaschi, L. Sala– *The "Digital Maktaba LP": Proposing a Comprehensive Dataset for Arabic Script OCR Title Pages in the Context of Digital Libraries and Religious Archives*
DOI: http://doi.org/10.60923/issn.2532-8816/22019

The "Digital Maktaba LP": Proposing a Comprehensive Dataset for Arabic Script OCR Title Pages in the Context of Digital Libraries and Religious Archives

Riccardo Amerigo Vigliermo

Fondazione per le scienze religiose, FSCIRE
vigliermo@fscire.it

Giovanni Sullutrone

Università di Modena e Reggio Emilia, UNIMORE
giovanni.sullutrone@unimore.it

Sonia Bergamaschi

Università di Modena e Reggio Emilia, UNIMORE
sonia.bergamaschi@unimore.it

Luca Sala

Università di Modena e Reggio Emilia, UNIMORE
luca.sala@unimore.it

Abstract

Optical Character Recognition (OCR) plays a vital role in digitising and enabling access to historical records in digital libraries. Yet, OCR technologies frequently face challenges when interpreting and categorising intricate document structures, particularly in historical materials with varied layouts and languages. This initial study tackles the issue by proposing the creation of a rich, publicly available dataset of Arabic title pages, leveraging advanced Vision Language

Models (VLMs) alongside OCR techniques. By extracting the first pages of each document in high resolution, our focus was on accurately identifying frontispieces and separating them from the main body of the text to enhance metadata accuracy and document discoveries within digital repositories. The Qwen-2vl-72B model was employed to determine whether each page is a frontispiece or non-frontispiece through a custom-designed prompt. Detected frontispieces will be processed with Google Vision AI to generate Ground Truth data, later reviewed by linguistic specialists before finalising the dataset. Future plans include training open source models like Kraken OCR to assess dataset utility. This novel strategy addresses current dataset gaps while boosting digital archive performance, including in projects like Digital Maktaba.

Keywords: Arabic OCR, Datasets, Title pages, Digital Libraries, Religious Archives

Il riconoscimento ottico dei caratteri (OCR) svolge un ruolo fondamentale nella digitalizzazione e nell'accesso ai documenti storici nelle biblioteche digitali. Tuttavia, le tecnologie OCR spesso incontrano difficoltà nell'interpretare e categorizzare le intricate strutture dei documenti, in particolare nei materiali storici con layout e lingue diverse. Questo studio iniziale affronta il problema proponendo la creazione di un ricco set di dati di pagine di titoli arabi disponibili al pubblico, sfruttando modelli linguistici di visione (VLM) avanzati insieme alle tecniche OCR. Estraendo le prime pagine di ogni documento in alta risoluzione, ci siamo concentrati sull'identificazione accurata dei frontespizi e sulla loro separazione dal testo principale per migliorare l'accuratezza dei metadati e la scoperta dei documenti negli archivi digitali. Il modello Qwen-2vl-72B è stato utilizzato per determinare se ogni pagina è un "frontespizio" o un "non frontespizio" attraverso un prompt personalizzato. I frontespizi rilevati saranno elaborati con l'intelligenza artificiale di Google Vision per generare i dati di Ground Truth, successivamente esaminati da specialisti linguistici prima di finalizzare il set di dati. I piani futuri prevedono l'addestramento di modelli open source come Kraken OCR per valutare l'utilità del set di dati. Questa nuova strategia affronta le attuali lacune del set di dati e aumenta le prestazioni degli archivi digitali, anche in progetti come Digital Maktaba.

Parole chiave: OCR alfabeto arabo , Dataset, Frontespizi, Biblioteche Digitali, Archivi Religiosi

Introduction

Digital libraries are those "conversations," or conversational spaces,[1] where a field like Digital Humanities (henceforth DH) finds its point of contact between technology and humanities in general. This is mainly because traditional libraries have been the closest Humanities field to informatics and technology, as remarked by Petrucciani in the preface to *Verso nuovi principi e*

[1] This defintion is provided by The *Nuovo manifesto per le biblioteche digitali*, inspired by the first *Manifesto per le biblioteche digitali* of the Association of Italian Librarians [5]. Its aim is to set the principles, models, and functions of digital libraries extending the previous *Manifesto* from 30 to 33 theses. The first principle is "Le biblioteche digitali sono conversazioni" (digital Libraries are conversations), where we read: "Non biblioteca digitale, ma biblioteche digitali, non un sistema, una grande narrazione sistematica, ma tante conversazioni tenute insieme da un linguaggio comune, da una struttura comunicativa basata sull'assunzione di impegni fra comunità diverse per pubblici diversi." ("Not digital library, but digital libraries, not a system, a great systematic narrative, but many conversations held together by a common language, by a communicative structure based on commitments between different communities for different audiences") [57].

nuovi codici di catalogazione ([58], 11): "The development of new information and communication media, new technological tools and new opportunities for reading and study has always accompanied that of libraries."[2] The same concept has been pointed out by Anderson ([18], 8) as a natural connection between computer science, information science, and library science with striking analogies that, once identified, have enabled (or should enable) the transition from managing cataloguing workflows to collaborative efforts in developing innovative tools within the DH. What expressed by Petrucciani and Anderson follows the same direction of what Burdick et al. [29] defined as the transformative and balanced approach between the critical-interpretive approach typical of the Humanities and an "empirical" [60] or "operational" one [32]. Librarians' deep understanding of metadata and information organisation principles becomes essential in designing robust digital systems. Several projects showed how effective this combination of expertise could be in creating sophisticated text analysis tools and digital platforms enabling cross-cultural digital heritage preservation and access. Those collaborations opened new research paradigms where the theoretical foundations of librarianship actively shape the development of digital tools, while computational methods enhance and expand traditional library practices.[3] Moreover, librarians, having a long tradition of exchange knowledge with informatics, could contribute significantly in reconsidering libraries as rich datasets to develop heterogeneous DH projects involving Machine Learning and AI techniques as part of the data librarianship field [86] and as proper metadata librarians [109], as well as contributing to the pedagogical role of the library in the digital age ([1], 111).[4]

An interesting starting point are those libraries and archives dealing with historical and religious materials, since religious studies methodology is inherently cross-disciplinary ([47], 147-148), representing, at the same time, a rich multi-linguistic, multi-alphabetic, and multi-confessional

[2] Authors' translation.

[3] We cite here just few examples, such as The Perseus Digital Library Project, which begun in 1985 and was focused on digitising and making accessible materials related to Greco-Roman literature, history, and culture from. This project maintains an open-source infrastructure and became an experimental platform for methodologies and a functional online repository of classical texts and resources, see http://www.perseus.tufts.edu/hopper/. Another example is the Europeana project, the European Union's digital platform for cultural heritage, launched in 2008. It serves as a vast online portal providing access to millions of digitised items including artworks, artifacts, books, photographs, music, maps, and archival materials from thousands of European museums, libraries, archives, and cultural institutions. The platform enables cross-border and multilingual search capabilities, promoting European cultural heritage while making it freely accessible for education, research, and creative purposes, see https://www.europeana.eu/it/about-us.

[4] The pedagogical aspect, while not the primary focus of this paper, is indeed important in avoiding what the same Adams recalls as "buttonology" from the definition given by Russel and Hensley [102]. The buttonology approach could be seen as the manualistic teaching (or learning in some cases) of a software or a tool or an interface, which is not the inherent goal of the DH research since it overshadows the critical space, ultimately preventing the possible comprehension of the Humanistic data in their more profound sense by only training the Humanists on how to perform certain tasks, without instructing them on the implications of the use of some software on their research questions.

dataset, making them an ideal testing ground for digital innovation ([18];[1]). For this reason, the study presented here has started from considerations on the extensive digital book collection of the Giorgio La Pira library in Palermo, part of the Fondazione per le scienze religiose (FSCIRE), dedicated to Islamic history and doctrines.[5] This collection forms the core dataset of the Digital Maktaba (henceforth DM)[6] project, which aims to develop methodologies for cataloguing and managing multilingual texts, enhancing librarians work, and facilitating access to historical-linguistic-religious knowledge. DM seeks to support the cataloguing of wide digital repositories in non-Latin scripts and provide scholars with advanced retrieval tools, while addressing historical-cultural considerations ([27];[83];[26];[82];[116];[122]). More specifically, DM is dedicated to crafting a digital library that can analyse and extract information from multi-lingual documents, particularly from Arabic scripts (Arabic, Persian and Azerbaijani), offering a state-of-the-art cataloguing methodology designed specifically for those libraries that need to manage multi-lingual and multi-alphabetic cultural resources, ultimately promoting inclusive library practices.

The general aim of DM and the multi-lingual and alphabetic nature of the data at disposal of DM inevitably crosses its path also with the growing demand for accessible and searchable digital texts. Optical Character Recognition (OCR) technology, which converts various document formats into editable and searchable data, became essential in this transformation. This technology is crucial for digital archiving, information retrieval, and data analysis satisfying the criterion of *functionality* over *exhaustivity* from an automatic processing of content perspective.[7] However, OCR technology still faces notable challenges, particularly with complex and historical documents. These challenges are amplified when processing Arabic script, where unique linguistic, typographic, and calligraphic characteristics contribute in affecting OCRs accuracy. When considering Arabic OCR in a cataloguing context, one of the primary barriers in OCR research and application is the lack of comprehensive, high-quality datasets designed for library usage. In comparison to other languages and scripts, existing datasets often lack the breadth and specificity required to address these complex features, hindering the development and benchmarking of advanced OCR algorithms. To provide some examples, Arabic

or Persian script datasets focus mostly on single modern handwritten characters ([97];[90];[12]), full text pages of content or text lines([100];[105]), subwords [107], isolated form characters or single characters ([120];[104]) or specific fonts ([10];[78]). At the same time, those datasets

[5] https://www.fscire.it/heritage/biblioteca-la-pira.

[6] Digital Maktaba (DM) is part of the Italian Strengthening of the European Research infrastructure Resilinece (ITSERR) project, which was launched in November 2022 and funded by the Italian Ministry of Research with NextGenerationEU programmefunds. It involves the University of Modena and Reggio Emilia, CNR, University of Palermo, University of Turin and the University of Naples "L'Orientale". Its objective is to improve the European Research Infrastructure RESILIENCE in response to the demands of the scientific community in Religious Studies in terms of technology integration and capacity to increase innovation.

[7] Functionality and exhaustivity were sharply analysed and discussed by Buzzetti ([31]:64) as two key concepts in digital libraries and preservation, especially of textual data (i.e., series of encoded characters). Exhaustivity is related to the of reproduction of any document into a digital representation. Functionality is instead related to the operational side or the possible analysis and elaboration of the textual components (i.e., the content) as represented in the digital form.

focused on historical text usually include only manuscript texts or early printed texts content ([84];[45]) as more relevant to the Humanistic research in comparison to title pages, which, however, are the essential source of cataloguing information in library science.

This limitation is especially critical for elements like frontispieces (i.e., title pages) which contain unique artistic, typographic, and calligraphic features that demand specialised handling. Figure 1 shows an example of title pages.

In developing a cataloguing tool, which is the main aim of the Digital Maktaba project, we present here a pipeline for the creation of a title pages dataset to effectively train an Open Source OCR model such as Kraken [72], through the eScriptorium VRE [113], to extract cataloguing metadata from Arabic printed frontispieces. The presented work also considers recent advances in Vision Language Models ([22];[75]) that could contribute significantly to data extraction from images by integrating visual, as well as textual understanding. By leveraging VLMs and a closed source OCR such as Google Vision AI this study addresses these challenges by developing an extensive, community-accessible dataset of title pages in Arabic script. We focus specifically on accurately classifying frontispieces and distinguishing them from the main text within the initial pages of each document. Additionally, we aim to generate enriched metadata for improved organisation and retrieval in digital libraries. Currently, we are in the process of constructing a dataset that captures the diverse typographic and structural challenges of Arabic texts. Our immediate goal is to finalise this dataset and utilise the best available OCR tools to automatically extract text, or portions of it, from these documents. This extracted text will then be meticulously reviewed and corrected by experts to create a gold-standard reference, ensuring accuracy for future OCR advancements. In the sections that follow, we provide background on OCR technology, elaborate on the unique challenges posed by Arabic script, outline our objectives and methodology, and discuss the preliminary workflow and the anticipated outcomes and implications of our study.

Background

Arabic script main feature with reference to calligraphic styles

The Arabic script consists of twenty-eight graphemes and has the following marking characteristics: it proceeds from right to left (RtL); it is essentially consonantal as it only transcribes consonants and long vowels (*ā*, *ī*, *ū*); consonantal tension (doubling) is not indicated in unvoiced texts; it is cursive because twenty-two of the twenty-eight graphemes link to each other within the word, changing their form according to context (isolated, initial position, middle, final). Short vowels (*a*, *i*, *u*) are marked with diacritics mostly in didactic texts, partly in poetry and fully in the printed versions of the Qur'ān. To be more specific, the diacritical marks corresponding to short vowels *a* is ـَ (*fatḥa*, e.g., لَ *la*), for *u* is ـُ (*ḍamma*, e.g., لُ *lu*), for *i* is ـِ (*kasra*, e.g., لِ *li*). To indicate that a given consonant is unvoiced, the symbol ـْ (*sukūn*, eg.لْ *l*) is placed on the grapheme. To indicate a double consonant instead, the diacritic ـّ (*šadda*, eg. لّ *ll*) is used. Those marks are considered as not part of the alphabet [43]. Arabic has regularised the use of *matres lectionis*, whereby long vowels (*ā*, *ī*, *ū*) are systematically written with consonantal signs respectively ا (*alif*), ي (*yā’*, *y*) e و (*wāw*, *w*). Its consonantal system richness led to the adoption of diacritical dots to distinguish so-called homograph consonants (different but of the same sign or spelling) such as: ر *r* e ز *z* or ض *ḍ* e ص *ṣ*. Homography is one of the main issues with Arabic script. Particularly noteworthy in Arabic is also the preservation of nominal inflection and the ability to develop features intrinsic to Semitic languages like: *triconsonantalism* of the root, the presence of pharyngeals and emphatics phonemes, and internal inflection, such as the internal plurals [52]. From a calligraphic standpoint, the earliest Old Arabic inscription (found in the South Arabian script, 328 AD) changed shape, over the course of two centuries, becoming more akin to Syriac and "Arabic" particularly in ligatures joined in straight lines on a baseline ([55], 21-22).

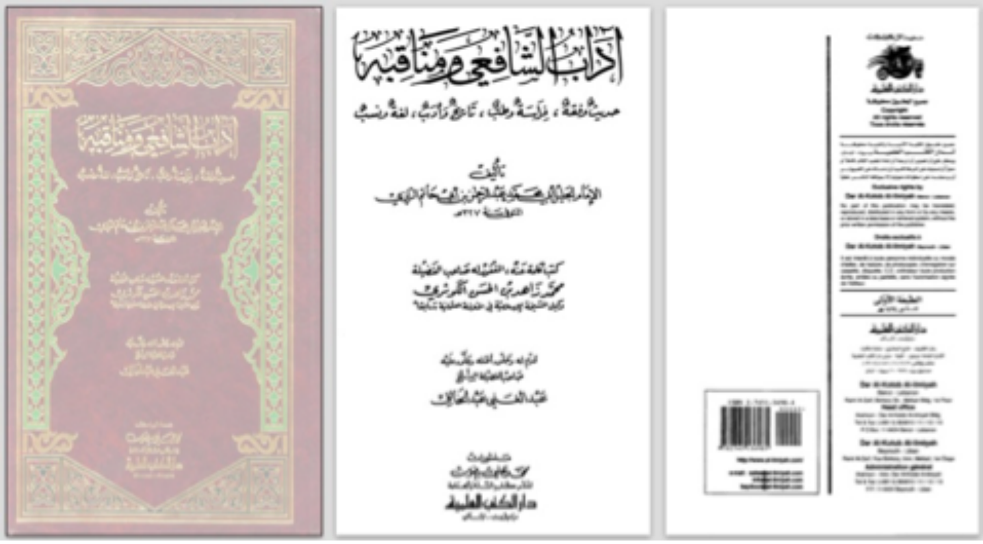

Figure 1 - Example of a Frontispiece Pages Group (FPG) as intended in this study

Even though sources on the origin of the Arabic script are varied and uncertain[8], several historical narrative examples seem to converge on three members of the Ṭaiyy’ tribe who gathered in the village of Baqqa and laid the foundation of the Arabic writing (*al-ḫaṭṭ*) by modelling the alphabet on the Syriac one (*qāsū al-hiǧā’ al-ʿarabiyya ʿalà al- hiǧā’ al-siryāniyya*). Then, the writing was learned by some people of al-’Anbār (*qawm min ’ahl al-’anbār*) that taught it to the people of al-Ḥīra (*’ahl al-ḥīra*). After that, an Arab Christian named Bišr ʿAbd al-Malik brought the writing learnt from al-Ḥīra to Mecca ([23], 759). The story has been reported repeatedly in different forms and lengths but with the same narrative by different sources such as: ([64], 4:240; [62], 11); who, for example, did not mention the Syriac *hiǧā’*; ([98], 3:12) who designated the three people with different roles in creating forms, ligatures, and diacritics. To resume, the

[8] This is true when looking at the two divergent research direction in the Arabic calligraphic studies: the first one locates the origins of *al-ḫaṭṭ* in the Himiaritc (southern Arabian) script (also called *musnad*), while the second one traces the origins of the Arabic script to Nabatean and Syriac-Aramaic ([115],33).

reported narrative points out a context where pre-Islamic Arabs and Arab Christians moving around the two Iraqi cities of al-'Anbār and al-Ḥīra and the Ḥiǧāz region of the Arabian Peninsula are involved. At the same time, the script seems to be first derived from the Nabatean script and then calligraphically influenced by the Syriac one ([55], 27). During the 6th century, the advent of Islam radically transformed the evolution of the Arabic script. Verses of the Qur'ān, transmitted orally and memorised, began to be written on leaves (*ṣuḥuf*) collected subsequently in codices *muṣḥaf* (pl. *maṣāḥif*)[9]. Primitive Arabic script was defined by different sources with the name of *ǧazm* ([62], 13; [63], 372), otherwise known as *ḥiǧāzī* (literally "from the Ḥiǧāz region")[10] according to Ibn Nadīm, who reports in his *Fihrist* the words of Muḥammad ibn 'Isḥaq, first is the Meccan style (*al-makkī*) followed by the Medinan one (*al-madanī*), the Baṣran style (*al-baṣrī*) and the Kūfan style (*al-kūfī*). The first two styles see the shape of the *alif* curved to the right and elongated upwards with a slight inclination with respect to the baseline.
With the expansion of the Arab-Islamic empire also calligraphy evolved and transformed. At an earlier stage, the *kufī* calligraphic style (from Kufa in Iraq) distinguished itself for clarity and quality. Under the Umayyad dynasty, calligraphy began to differentiate. Masters like "Quṭba" and Hasan al-Basri invented styles such as *al-ṭumār* and *al-ǧalīl* (nowadays *al-ǧālī*). During the Abbasid period, figures like al-Ḍaḥḥāk bin 'Aǧlān e Isḥaq bin Ḥamād became prominent for their craftsmanship in the *al-ǧalīl* style. During the same period new calligraphic styles such as *al-siǧǧilāt, al-dībāǧ, al-ṯulṯayn, al-ṭumār al-kabīr* e *al-ʿuhūd* emerged. Under al-Ma'mūn, the process of innovation continued resulting in the development of new styles: *al-muraṣṣaʿ*, *al-nassāḫ*, *al-riyāsī*, *al-riqāʿ* e *al-ġubār*. The main calligrapher of this period is considered Ibrāhīm al-Šaǧarī who created the *ṯulṯayn* ("two-thirds") style and the *ṯuluṯ* ("one-third") style, just to name a few.[11] Ibn Muqla, Ibn al-Bawāb and Yāqūt al-Muʿtaṣimī followed him and from the 7th century new calligraphic styles emerged from the contacts with non-Arab Islamised peoples. This context gave birth to the *taʿlīq* style thanks to the mastery of Turkish and Persian calligraphers in the 8th century, consolidated and being followed by its variant *nastaʿlīq* invented by the persian master Mir ʿAlī Tabrīzī and particularly popular in Central Asia ([108], 32-34). Despite the multitude of calligraphic styles produced, only a few are still commonly used today in particular publications of both classical and religious works. Among the styles employed are: *kūfī, muḥaqqaq* and *rayḥānī, nasḫ, ṯuluṯ, diwānī* and *diwānī ǧālī, ruqʿa, taʿlīq, nastaʿlīq, šekasteh.*[12] In

[9] From the root *ṣ-ḥ-f* , Pass. participle of the IV verb form *'uṣḥif*^{a}*-yuṣḥaf*^{u}: "written pieces of paper or of skin collected in it, or put in it between two covers" see for example *muṣḥaf* in Ġarīd al-Šayḫ .

[10] For further details on this primitive Arabic script such as the form and the presence of medial *alif*, *tā' marbūṭa*, wāw, ṣād and so on, see George ([55], 32-33).

[11] For other styles see, al-Qalqašandī, *Ṣubḥ al-'aʿšà*, III, pp. 13-6. [98]

[12] In terms of subdivision, studies on calligraphy do not provide an unambiguous subdivision of the various calligraphic styles. Some scholars, for instance, consider only six calligraphic styles to be recognised as those of the calligraphic tradition (*ṯuluṯ*, *muḥaqqaq*, *rayḥānī*, *nasḫ*, *tawqīʿ* and *ruqʿa*). See for example Safadi ([106], 17) ([106], 17). Al-'Alūsī analysed more deeply some styles (*kufī ṯuluṯ*, *muḥaqqaq*, *rayḥānī*, *nasḫ*, *ruqʿa*, diwānī and

addition to these, there are other less common styles, such as *siyāqa, sunbulī* and *ṭuġrā* (or *ṭuġrāʾ*).[13]

Ṣubḥī Murād identifies two additional calligraphic styles that are particularly relevant to this research: *ḥadīṯ* (“modern”) and *ḥurr* (“free”). The modern style originates from the *kūfī* script but adapting its traditional features to a more creative framework without always adhering to the graphical relationships between graphemes. The *ḥurr* style breaks away entirely from traditional constraints ([115], 445-450). Graphemes, ligatures, and diacritical marks result distorted, curved, fragmented, elongated, or slanted to the limits of readability.[14] This approach prioritises aesthetic balance over textual uniformity, making it particularly suited for advertising, publishing, and contemporary artistic expression. Both styles have played a key role in the design of book covers, journal titles, and newspaper mastheads. This evolution is significant in the context of OCR applications in librarianship, where the artistic fluidity of the “free” style presents unique challenges for automatic title and author extraction.

Arabic Optical Character Recognition in Digital Libraries: The title page as a Group and a Visually Rich Document

OCR technology has played a key role in library and archive digitisation efforts across the globe. By converting printed text into machine-readable formats, OCR facilitates the storage, retrieval, and analysis of vast document collections. Traditional OCR systems rely on pattern recognition and machine learning techniques to interpret character shapes and word patterns in scanned images [93]. However, OCR systems are still lacking in training on non-digital native texts, particularly in the historical or religious domains. Large amounts of unstructured data with dimensional, dispersion, diversity, and noise characteristics expose the limitations of information extraction techniques on several levels: nature of the data, their usability, language and domain limitations, capacity of the techniques and approaches used [2]. The limitations, as will be seen shortly, become even more evident when the characters processed are Arabic characters [4]. What has just been said provides an indication of the current state of OCR on the Arabic alphabet. Often, the outputs generated by OCR systems are inaccurate due to several variables, which represent a real problem if one considers OCR systems as central to the development of

diwānī ǧālī ,etc.), than others (*tawqīʿ*, *siyāqat*, *sunbulī* etc.), see [16]. Other scholars divided between older styles and present styles (*ṯuluṯ*, *nasḫ*, *tawqīʿ*, *ruqʿa*, *diwānī*, al-*fārsī*, *siyāqat*, *kufī*, *maġribī*, *rayḥānī*, etc.), see [11] and [115]. On the contrary, other studies do not pay much attention on the subdivision itself but on the artistic and mystic value of every single style, as in the case of Schimmel and Rivolta [108]. The purpose behind the division here (main and other calligraphic styles) is only to report the most known and widespread Arabic script styles that are actually present in many title pages, so particularly interesting in the context of automatic Arabic text extraction.

[13] For more details on the *sunbulī*, *dīwānī* and *siyāqa* styles, see Faḍāʾilī ([48], 418); for the *ṭuġrā* see al-Alūsī ([16], 59).

[14] An interesting work on the *ḥurr* style has been written by ʿAfīfī [130]. The author gathered insights from modern Arabic calligraphers, artists, designers, and publicists, revealing that the *ḥurr* style follows a cyclical pattern of innovation and rediscovery, while being still rooted in the original *rasm*..

applications involving the successive use of other Natural Language Processing (NLP) techniques, which use the outputs of the former as their input even for very different tasks ([35];[114]). In this sense, recent studies have indicated promising areas of research focusing in particular on the pre- and post-processing phases. Post-processing techniques for OCR system outputs based on semi-automated approaches are already being developed by following error models and implementing sophisticated mathematical, linguistic, and probabilistic solutions in order to correct outputs ([95], 9). Other significant approaches see context-based solutions through the use of spellchecking systems [24], statistical translation machines [3], or rule-based solutions [70]. Even in the field of studies on the digitisation and preservation of vast corpora of historical texts, the importance of misread outputs analysis is not underestimated; on the contrary, it is placed at the centre of the agendas of digitisation projects of multilingual text resources [112].

If what has been said could be valid and effective for documents with conventional fonts and layouts, these systems struggle with documents that deviate from these standards. Moreover, the Arabic script poses several specific challenges due to its unique linguistic and typographic features. Arabic cursive, RtL script, complicates character segmentation [49]. Some Arabic graphemes have diacritical dots (above or below the baseline) that can shift in different calligraphic-typographic styles leading to recognition errors. The script primarily represents consonants, with vowels indicated by optional diacritical marks (that are often not presented). Arabic characters change shape depending on their position within a word. Some graphemes do not connect (bind) with others creating words with multiple disconnected components. Lastly, in some calligraphic styles the characters may overlap, touch, or appear in slanted orientations. All these graphical challenges conduct us to consider the frontispiece as only a part of a series of pages that we refer to as Frontispiece Pages Group (FPG). This definition is motivated by the context of frontispiece OCR analysis and characters extraction for the development of a system able to support the librarian work. As shown in Figure 1, FPGs could be conceived as a group of pages where most of the metadata useful for cataloguing is present: it is a group because we often have re-propositions of the same information in different scripts (e.g. title is represented in the title page, as well as in other pages, sometimes using other fonts for the sake of simplifying the cataloguing process). The title page in many cases is a black-on-white re-proposition of the cover page where the binarisation of the text does not solve other graphical issues (calligraphic peculiarities, decorations, vocalisation, etc.). Subsequent pages of the FPG usually reports useful data (even in a fragmented manner across several pages) in more normalised scripts enabling an easier text extraction and cataloguing. In some cases, the information is placed inside special boxes on one of the pages following the title page. From a Document Analysis perspective, the title page could be considered the same as a Visually rich Document (VrD), especially considering that scanned PDFs from physical realm could bear also noise elements such as libraries stamps or marks and several other issues related to the state of the paper as support. VrD is a term used in document analysis mainly for business or daily life documents (e.g., purchase receipts, insurance policy documents, custom declaration forms, and so on). Several studies on VrDs have emerged recently proposing graph convolution based model [76], a dataset [123] più sotto, and recent document and layout analysis studies [94].

Peculiar challenges posed by Arabic script FPGs

Frontispieces present unique challenges for OCR extraction that are not typically encountered with internal pages, as they often exhibit a high degree of variability and complexity. Several factors contribute to the increased difficulty:

Variety in Layouts and Designs. Frontispieces may include ornate designs, decorative elements, and unconventional layouts that intertwine text and images. This visual richness can confuse OCR systems, which are primarily trained on text-centric pages.

Diverse Backgrounds and Noise. The presence of backgrounds with various colours, textures, or deteriorated conditions adds noise to the images. Such backgrounds can interfere with text recognition by obscuring characters or creating false positives.

Non-Standard Fonts and Scripts. Frontispieces often feature artistic or custom typefaces, including calligraphic styles like *Naskh*, *Nastaʿlīq*, or *Kūfī*. These fonts have unique graphical peculiarities that are not always well-represented in standard OCR training datasets.

Multiscript Content. They may contain text in multiple scripts, such as Arabic and Latin, sometimes within the same page. Moreover, the use of numerals (e.g., for dates) and alphabetic script is by itself a challenge since Arabic-Indic numerals have a Left-to-Right (LtR) orientation while Arabic script follows a Right-to-Left orientation. Consider for example the string سنة ١٤٢٠ that could represent a publication date, or an author death date on a title page. In this case the RtL orientation of the Arabic word *sana* ("year") is opposite to that of the numerals "1420" causing an OCR system not adequately trained in segmentation, feature extraction, and classification to provide an incorrect Unicode mapping and representation of the text. In the case of an author death date, an incorrect recognition may lead to complication in retrieve such data which is an important disambiguation element in statement of authority record. This is just an example; however, it must be noted that the Unicode standard itself has ambiguous calligraphic styles. For example, the character ی has two codes – U0649 for Arabic and U06CC for Persian – despite being visually identical. Similarly, ه (<h>) is encoded as U0647 and U06D5, with variations in usage across languages like Azeri Turkish. These homographic conflicts impact precise cataloguing, as seen in the identical forms of *alif maqṣūra* in Arabic and *eżāfe* in Persian, which have different linguistic meanings ([67], 5).

Presence of Diacritical Signs and Marks. The inclusion or omission of vowels and diacritical marks can vary, affecting character recognition. Decorations or artistic elements might be mistaken for diacritics, leading to misinterpretation of the text. These challenges are compounded by external variables such as overall image quality, character resolution, different levels of support degradation and the presence of coloured fonts or backgrounds. Previous works ([27];[26];[83];[82]) highlighted that these issues require not only advanced OCR algorithms but also carefully curated datasets.

As we will explain further in the following sections, the proposed pipeline combines the Vision Language Model (Qwen-2VL-72B) with targeted prompt engineering. The pipeline is able to recognize the *Frontispiece Pages Group* (FPG) as a structured unit, rather than treating the title page in isolation. This strategy mitigates the impact of decorative layouts and heterogeneous typography by exploiting recurring visual cues and paratextual patterns across initial pages. Once identified, these pages undergo OCR via Google Vision AI, which supports right-to-left recognition and returns bounding geometries, thereby enabling both accurate text extraction and the structuring of metadata. In this way, the workflow not only reduces the high error rates typically introduced by Arabic homography, Unicode inconsistencies, and complex visual ornamentation, but also generates a "silver-standard" dataset of enriched cataloguing metadata that can be iteratively refined into a gold-standard resource for training open-source engines like Kraken. This layered approach effectively transforms the Arabic-script title page—from one of the most error-prone loci in digital cataloguing—into a reliable entry point for bibliographic metadata enrichment.

Related Works

Arabic script processing

Research on the automatic recognition of handwritten Arabic and Persian characters has evolved from early statistical and Hidden Markov Model (HMM) approaches to today's deep learning systems. In the 2000s, studies focused on artificial neural networks, HMMs, segmentation, and preprocessing ([77];[13];[50]). Hybrid methods combined contour analysis, probabilistic models, and HMMs with normalization and skeletonization ([68];[9]). Continuous HMMs [42] and planar HMMs [121] improved complex ligatures, while IFN/ENIT (a standard Arabic handwriting benchmark) became central. By the 2010s, deep architectures outperformed traditional methods: Multidimensional Recurrent Neural Networks (MDRNNs) with Connectionist Temporal Classification (CTC) [56], Bidirectional Long Short-Term Memory (BLSTM) networks [59], Convolutional Neural Networks (CNNs) ([44];[79]), and later VGG, ResNet, and transformer-based models [87]. Recent advances include Generative Adversarial Network (GAN) augmentation [91], Multi-Dimensional Long Short-Term Memory (MDLSTM) with Maxout [80], and hybrid CNN-Recurrent Neural Network (RNN) approaches for manuscript layout analysis [7]. Recognition rates now exceed 99% on datasets such as AHCD (Arabic Handwritten Characters Dataset), HACDB (Handwritten Arabic Characters Database), and KHATT, though challenges remain for noisy historical manuscripts. Persian OCR followed a parallel path, though with fewer studies. Early work used Zernike moments and HMMs ([38];[39]), fuzzy logic [21], and wavelet/fractal features ([88];[89]). Recent efforts combine CNNs with Error Correcting Output Codes (ECOC), Support Vector Machines (SVMs), and RNNs ([66];[19]), reaching ~96% with Hoda, IBN SINA, and the Bina system ([71]). Printed OCR progressed from contour/morphology ([30];[119]) to segmentation-free HMMs [69], Scale-Invariant Feature Transform (SIFT) features [127], and deep CNN-RNN hybrids ([51];[85]). Recent work explores You Only Look Once (YOLO) detection and transformers ([8];[87]). Today, Arabic and Persian OCR sits at the AI–humanities crossroads, enabling large-scale digitisation while open problems persist for multilingual, historical, and manuscript settings.

Vision-Language Models in an OCR pipeline

In very recent times, the focus on the multimodality of Large Language Models brought to a consistent and continuous release of several Vision-Language Models such as OpenAI's CLIP

[99] and GPT-4V [96], Salesforce's BLIP [73], DeepMind's Flamingo [6], and Google's Gemini [54].

Vision-language correlation examines how training goals and architectural strategies enable effective multimodal integration. The design of vision-language architectures shapes how well models can merge visual and textual information. Earlier models were typically trained from the ground up, while newer approaches are built on pre-trained language models to strengthen visual understanding through enhanced vision-language alignment ([74], 1-2). Those constantly growing models in terms of users interaction and multimodality found applications across a range of fields, including image captioning, visual question answering, and object recognition, where both visual and textual data are combined [34]. Leveraging large-scale neural networks and extensive datasets, VLMs are designed to interpret complex images that may include text, graphics, and other visual elements [75]. Although VLMs are still emerging in document analysis, their potential for handling mixed-media layouts and intricate document structures is promising. When considering VLMs model in comparison to traditional OCR systems, focused solely on recognising text, the former take a holistic approach by analysing both visual and textual components of an image. For instance, the Qwen-2vl-72B [22] model is designed to analyse multimodal data and can perform tasks such as image captioning and visual context interpretation, which could theoretically aid in recognising text within complex visual contexts. VLMs such as GPT-4V and Gemini, for example, have also been investigated for different text-related tasks including recognition, Scene Text-Centric Visual Question Answering (VQA) (Tang et al. 2024), Document-Oriented VQA [40], Key Information Extraction [125], and Handwritten Mathematical Expression Recognition [28]. As recently pointed out by Z. Li et al. ([74], 5), the number of VLM benchmarks has grown rapidly with the quick development of those models. At the present moment nearly 54 vision-language benchmarks for evaluation, including text-image, are present. At the time of writing, Qwen is one of the best overall performing open-source VLM according to benchmark evaluations [124].

Arabic Printed Characters Datasets

In the last two decades Arabic script OCR studies have made significant steps forward. One of the first databases for printed Arabic characters was created in 1995 by the Environmental Research Institute of Michigan and therefore called ERIM. Comprising a set of Arabic books and journals, it contains about 750 pages with approximately one million characters and more than 200 bindings. The pages were scanned at a resolution of 300 dpi and the database was divided into three separate sets: training, statistics, and tests. One disadvantage of ERIM is that it does not include different aspects of written communication than magazines and books; letters, newspapers and so on are not covered. The second disadvantage is that it is not available online for free. Subsequently, Davidson e Hopely [37] started DARPA (Defense Advanced Research Project Agency) on request of the US Department of Defence. The database contains 345 images with a total of approximately 670,000 characters. The images, scanned at a resolution of 600 dpi, contain columns in which the text varies in quality. The corpus was obtained from books, journal articles, magazines, and computer documents in four different fonts. However, like ERIM, the database does not cover all aspects of writing and is not available free of charge due to its nature.

An example of database useful for both handwritten and printed (offline or online) Arabic text is ARABSE [17]. The database is a composite who collects images of documents, PAWs, isolated characters, digits, free text, and so on. ARABASE is a relational database that contains also an Arabic writing recognition system, an interface to experiment different image processing tasks. In 2009, the work of Slimane et al. led to the publication of the most famous and widely used database in Arabic printed character recognition research: APTI (Arabic Printed Text Image). Composed of 45,313,600 images covering 250 million characters, it is one of the most extensive databases. Its numbers are the result of the variation of: ten fonts, ten sizes, and four styles respectively of 113,284 words in .xml files. The APTI database is synthetic and designed for the evaluation of OCR systems by means of an automatic programme that generates images at 72 dpi, which can be disadvantageous, however, when scanned print documents are to be evaluated. A year later, Al-Hashim e Mahmoud [10] have developed a new database for research on printed Arabic character recognition. PATDB (Printed Arabic Text Data Base) is a corpus consisting of several scanned images of Arabic texts of various kinds at different resolutions (200, 300, and 600 dpi). 6,954 pages were collected in order to make them available for research and to make the database the reference for algorithm evaluations and comparison of results obtained by testing on the same database. PATDB is the first valid example of a database that is completely free and available in its entirety on the net. Al-Muhtaseb [14], drawing form the APTI database, published the Printed Arabic Text Set A01 and A02 (PATS-A01, PATS-A02) consisting of 2,766 text line images. The text of 2,751 line images of this set were selected from two standard classic Arabic books; that of the remaining 15 lines images is taken from minimal Arabic script. The line images are available in eight fonts: Arial, Tahoma, Akhbar, Thuluth, Naskh, Simplified Arabic, Andalus, and Traditional Arabic. The individual text lines of the PATS-A01 database were segmented manually to separate them into words to generate 24 training classes (13 classes for PATS-A01 and 11 class for APTI) for different Arabic words in different sizes, orientations, noise degrees, and fonts.

A further attempt is the work of Jaiem et al. [65] which developed a multi-font database for the evaluation of recognition systems using open vocabulary. APTID/MF (Arabic Printed Text Image Database/Multi-Font) is designed for segmentation research and automatic font identification and consists of 387 pages of documents scanned in greyscale at 300 dpi resolution. From each document, 1,845 text-blocks were extracted with their original files; in addition to the text-blocks, a large dataset of 27,402 samples is provided. This database, like the previous two, is available to researchers free of charge. Also of similar design was the ALTID database, useful for both printed and manuscript characters for both the Arabic and Latin alphabets. The printed text fragments (1,845 in Arabic and 2,328 in Latin alphabet) were obtained by manual segmentation from 731 greyscale images with a resolution of 300 dpi [36]. As far as the recognition of Arabic fonts is concerned, mention is made here of the KAFD database [78]. The database consists of 40 fonts in 10 sizes (8 to 24 pt.) and in 4 different styles (normal, bold, italic, and bold cursive). It is divided into three sets: training, validation, and testing respectively, and is open source for researchers. An example of bilingual database is LAMIS-MHD. A database designed for signature verification, writer recognition and writer demographics classification and also isolated digit recognition and similar related tasks. The database comprises 600 Arabic and 600 French text samples, 1,300 signatures and 21,000 digits. 100 Algerian individuals coming from different age groups and educational backgrounds contributed to the development of database by providing a total of 1,300 forms [41]. A public database for Arabic text images included in TV broadcasts is ALIF [126]. Its dataset consists of a collection of manually annotated text images and represents the first dataset dedicated to the development and evaluation of OCR systems in a video context. It is important to emphasise that text images have a great variability of fonts, size, colours, and quality. In the same context can be found the AcTiv database, designed for the identification of Arabic text within videos. For this purpose, 80 videos (850,000 frames) from four Arabic news channels were collected in an attempt to cover the maximum diversity of position, size, content, and background of the text-boxes. Each text-box is annotated in detail and presented with a region-based approach and a set of evaluation protocols for measuring performance [128]. Saddami, Munadi, and Arnia [103] published a 1,524 printed Jawi characters database from four types of fonts. The database also includes 168 printed word and sentence images. SmartATID [33] is dedicated to images of Arabic texts captured with mobile devices such as smartphones. The processed images can be useful for numerous tasks: font recognition, author recognition, word or line segmentation. Bataineh [25], in his study, presented a new database concept by collecting sub-words (Part of Arabic Words, PAW) instead of words or individual characters. The 83,056 PAW images are collected from approximately 550,000 different words. Each sample is presented in five different fonts: Thuluth, Nasḫ, Andalusi, Typing Machine, and Kufi for a combined total of 415,280 images associated with a statistical analysis of the frequency of each PAW in the Arabic language. In 2018, an extension to the AcTiv database was published: AcTiv 2.0 [129]. The new dataset, dedicated to the creation and evaluation of Arabic text recognition systems in video, contains:189 videos, 4,063 keyframes and 10,415 text images. As with its predecessor, each element is distributed with relevant annotations and open-source evaluation tools. Finally, two datasets were designed and developed for the recognition of Arabic printed text with examples and text images collected from the Qur'ān: QTID (Quran Text Image Dataset) consisting of 309,720 images with a total of 2,494,428 characters from the Qur' ān ic text [20]; and the second [15] containing 604 images at page level and 8,927 images at text line level from the Medina Qur'ān (*muṣḥaf al-madīna*). At the best of our knowledge no specific dataset has been developed for the analysis of Arabic printed title pages OCR handling in the context of librarian use and digital libraries.

Methodology

The primary focus of this study is to improve the OCR capabilities for digital libraries by developing a comprehensive, community-accessible title pages dataset. Our methodology involves assembling an initial set of historical documents from the FSCIRE La Pira library digital archive, classifying their pages as "frontispiece" or "non-frontispiece", and creating an initial OCR draft of the former using Google Vision AI, which will be evaluated with common metrics such as Character Error Recognition (CER) and Word Error Recognition (WER), analysed and corrected by linguistic experts. The use of a system such as Vision AI allows us to balance the need to produce a sufficient number of examples without human intervention (approximately 70,000 examples excluding a part of the first batch used for the creation of the *golden-standard*) with the highest possible quality in order to then train an open-source OCR model within the context of a tool such as DM. The development of a dataset of title pages, once completed and published in the final version, will allow other libraries and other organisations to train other models on Arabic layouts and characters with a particular focus on printed title pages. This choice

was also motivated previous tests of open-source solutions such as Easy OCR, Tesseract, and Google Docs [26] where Vision AI emerged for better performances on the languages in our collection. This is further confirmed from other studies that evaluated Google Vision AI against other OCR engines such as the same Tesseract and Textract [61].
Despite that, the large number of images to be processed and the accuracy and flexibility needed for our specific goal led us to the selection of Google Vision AI. A visual representation of the proposed pipeline is shown here in Figure 2.

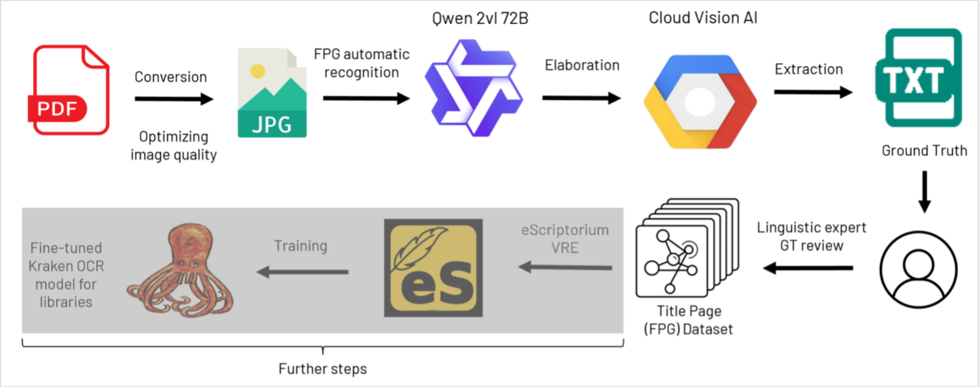

Figure 2 - the proposed pipeline to build the Digital Maktaba LP dataset. In grey some further step of training and fine tuning a Kraken OCR

Document selection, preparation and management. As anticipated, the proposed pipeline focuses on the large digital collection made available by the La Pira Library and which is composed by a heterogeneous donation of digital documents on Islamic studies, aggregated from different donors representing different institution from Iran to Jordan.[15] From a numerical standpoint the collection is composed of nearly 200,000 files, at least 70% of which are PDFs, for a total size of nearly 1.2 terabytes. We collected approximately 140,000 donated PDF documents, with high diversity in content, formats and cultural significance. This collection ensures that the dataset spans a range of subjects, periods, languages, layouts, fonts, and visual elements, aligning with our goals of cataloguing and preserving large non-Latin cultural heritages. To manage resources effectively and maintain consistency, we extracted the first ten pages of each document at high resolution labelling them as head (e.g., document name_head.pdf), as these initial pages typically

[15] The first institution in question is the Specialised Library on Islam and Iran of Qom pertaining to the University of Religions and Denominations. This notable collection covers several topics of religions and denominations in several languages in Iran ranging from Islam, Christianity, Judaism, Hinduism, Buddhism and primitive religions. The second is the Prince Ghazi Trust for Qur'ānic Thought and its Qur'ānic thought site which is a *sunnī šāfʿī* religious trust (*waqf*) and a very distinguished project aiming to collect and provide access to all the important texts on Islamic sciences, from ancient times until nowadays in a neutral, non-political, and non-fanatic way, promoting the consciousness and the knowledge of all aspects of the Islamic studies tradition.

contain frontispieces and other introductory materials. Each page is scaled proportionally to 512 x 512 pixels to reduce the number of converted tokens for the VLM used in the next step. The selection of the fist ten pages is crucial in reducing hardware and storage needs by limiting the processing on the most likely FPG pages

.

At the same time the VLM classification of pages helps to focus the OCR effort on the most informative pages from a cataloguing perspective. Reducing the time and resource-consuming effort will consent also bring forth the "AI in the loop, human(ist) in charge" paradigm behind the pipeline, which will include human expertise to linguistically review the OCR extraction to define a high-quality Ground truth dataset. The curated dataset will be tested and employed internally by training a Kraken OCR model for Arabic title pages for the purposes of the DM projects; once tested, the final aim is to publish the work as an open access, open source dataset for the training and fine-tuning of models specifically designed for libraries, as well as for other research fields.

Processing documents and preliminary results. To process these pages we selected the open source model Qwen-2vl-72B. Using a specialised prompt, we guided the model in classifying pages as either frontispieces or not frontispiece. The specialised prompt subsequently focused on the classification of different features present in the book, starting from the title page, intended as the preferred source of information, building upon the International Standard for Bibliographic Description (ISBD) at its last update ([46], 17-18). From the ISBD the title page could be replaced by a substitute such as the cover or the spine or other pages bearing the same information. The redundancy of information is one of the key elements of our FPG paradigm, as described previously.[16] It is worthy to remind that the FPG proposal described here is tailored on the challenges of Arabic script in library science and the digital environment, as well as being organic to the development of the Arabic OCR discourse and the application of this technology in support to librarians' work. Going back to the prompt, the frontispiece section (consisting in the group of pages defined as FPG) was also extended to pages containing relevant information for the cataloguing process and were most likely to be found in a book, namely, the ISBN code and the index. Since the latter two are less relevant to the present study we will focus here only on title pages classification. Hereafter an extract from our prompt:

> *prompts = {*
>
> *"frontispiece_classification": (*
> *"The preferred source of information is the title page (Frontispiece), or, for resources lacking a title page, the title-page substitute. If information traditionally given on the title page is given on facing pages, with or without repetition, the two pages are treated as the preferred source of information. The title page (Frontispiece) is a page normally placed at the beginning of a printed resource (usually black and white), presenting the most complete information about the resource and the works it contains [...] frontispiece is typically the decorative or illustrated page at the beginning of the book, often containing the title, author's name, and possibly publisher information or decorative elements. Other pages close to the frontispiece have usually useful information*

[16] It is interesting to note that the same section regarding printed resources does not mention the title page for non-Latin (non-roman) scripts (ISBD, A.4.2.1.2) but employs the term "colophon," as those sources in non-Latin script were only related to manuscripts or at least early-print material. However, standards were not specifically designed to address the peculiar challenges of the Arabic script and the calligraphic features present on the title pages.

> *and must be checked carefully [...] In some cases close to the title page you will find a page containing a CIP record (could be in a highlighted textbox or in a two column form). In this CIP there are the book's title ('عنوان الكتاب'), subtitle, and author(s) ('المؤلف' or 'تأليف'); ISBN code; Subject classifications (such as Dewey Decimal or Library of Congress Classification numbers etc.); Descriptive information like pagination, dimensions, and illustrations; Publication details (e.g., publisher and date).*
>
> *If any of those elements are present in any page close to the title page then consider it as 'Frontispiece' otherwise you should classify those pages as 'Not Frontispiece'. Only if the page is empty consider it as 'Not Frontispiece'*
>
> *Focus on:\n"*
>
> *"- Presence of distinctive Arabic typography or calligraphy for the title\n"*
>
> *"- Decorative borders, Islamic geometric patterns, or ornamental designs\n"*
>
> *"- Publisher logos or imprints\n"*
>
> *"- Author information placement\n"*
>
> *"- Barcodes \n"*
>
> *"- Body of text, which presents more lines of text and more characters, must not be considered as 'Frontispiece'\n\n"*
>
> *"Respond with either 'Frontispiece' or 'Not Frontispiece' [...]")*
>
> *}*

For the preliminary evaluation, the VLM was tested on a sample of 100 documents selected from both Arabic and Persian materials. While the sampling process was random, it was guided by an expert in order to maximize variation and avoid redundancy across the set. Specifically, documents were drawn from different sub-collections to ensure exposure to diverse calligraphic styles, title page layouts, and typographic conventions. This approach sought to balance the efficiency of random sampling with a deliberate effort to capture heterogeneity within the corpus. We acknowledge, however, that a sample of 100 items cannot be considered fully representative of the 140,000-document collection, particularly given the high variability of title page design. Accordingly, this evaluation is presented as a snippet of an ongoing work to expand the sample size using stratified sampling methods in order to better capture the full range of layout variation. Every page image has been processed separately one at a time, with the specific caveat of classifying multiple pages as "Frontispiece" in order to maintain the information redundancy, on the one hand, and extract as much cataloguing information as possible, on the other. The outputs on this first subset have been evaluated manually by a human expert and Qwen reached the promising performance with a Precision of 92%, a Recall of 93.5 % and a F1-score of 93%, according to the prompt given as a *zero-shot* for every instance. If we consider also the presence of some false positives (e.g., pages that have similar layouts and could resemble a title page but are not related to it) the performance dropped to 90%. However, the pages misclassified as title pages are in most cases bearing useful information on the author, the edition, or other data. The following figures (3 and 4) shed some light on those pages considered as correct outputs of a title page and false positives.

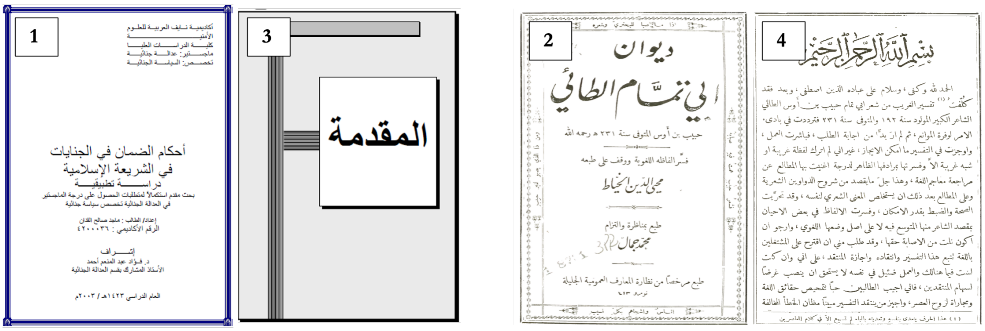

Figure 3 – Frontispieces (1-2) vs False Positives (3-4)

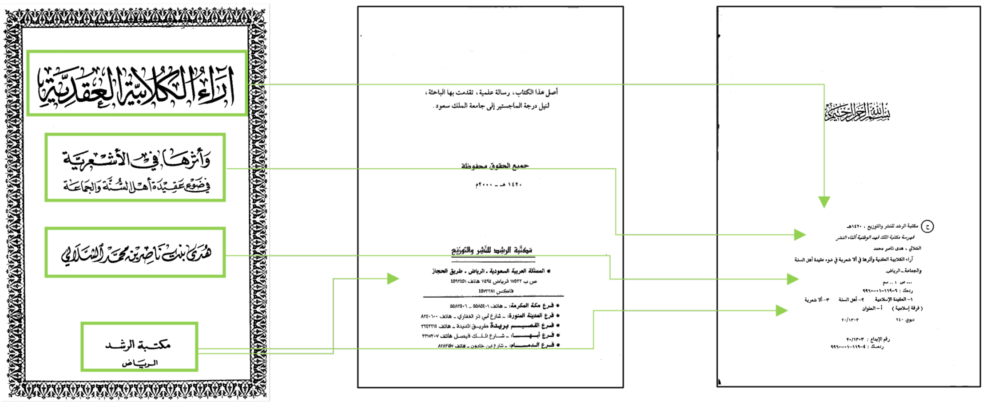

Figure 4 – An example of title page and neighbouring pages composing the FPG classified correctly

As reported in the two panes, images 1 and 2 have been recognised as "Frontispiece" correctly while the pages on the right in both panes (3-4) have been classified as "Frontispiece" probably on a layout basis. Image 3 in the left pane is reasonably misclassified since the layout is similar to that of a book title rather than a section title, which is what it is (i.e., the text is composed by the word *al-muqaddima* in Arabic meaning "the introduction"). However, outputs such as image 4, where the title recites the *basmala* (*bi-ism Allah al-raḥman al-raḥīm*, "In the name of God the Merciful, the Compassionate")[17] and a content layout is displayed, also are misclassified notwithstanding the specific command in the prompt: "Body of text, which presents more lines of text and more characters, must not be considered as 'Frontispiece'" . Those cases, though limited, seems to be less likely to occur on layout basis and more on some prompt misinterpretation from the model.

[17] This is one of the most important phrases in Islam and it is the opening phrase of all the chapters of the Qur'ān as well as introducing religious practices, writings, and daily activities.

Figure 4 and 5 show how the VLM selected the correct pages (in this case head_page_1, head_page_2, head page_4 from left to right) among 10 different pages in input. White and other less relevant pages are correctly not labelled as "Frontispiece". This case illustrates the concept of FPG as defined in this study. The image on the right presents the same data as the title page but in a simplified form—without decorative elements or diacritics. These simplified characters are repeated across adjacent pages, forming an information group that contains the complete cataloguing details (as indicated by the arrows in Figure 4). The normalised version of the title page characters is usually a simple *nasẖ* character easy to read and without any decoration or dicartical marks (vowels) to consent an easy way to retrieve cataloguing metadata. From an OCR perspective, this redundancy is valuable: it allows for identifying the most normalised representation of the title page text and using it as a sort of an "internal ground truth" to extract and link other graphic representations of the same string, word, or even individual character.

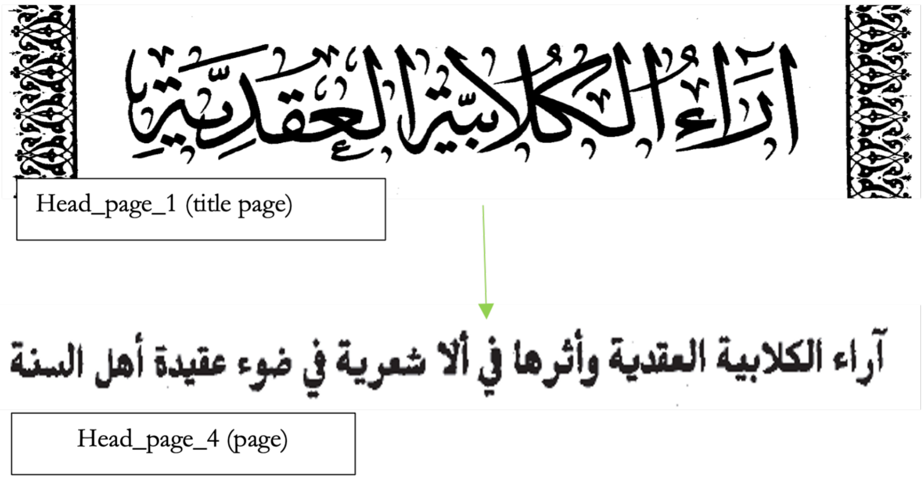
آراء الكلابية العقدية
Head_page_1 (title page)
آراء الكلابية العقدية وأثرها في ألا شعرية في ضوء عقيدة أهل السنة
Head_page_4 (page)

Figure 5 – Title on the title page and its "normalised" representation on another FPG page.

At the time of writing we are expanding the tests on a wider subset before extending it to around 20,900 pages of several FPGs. Once classified, those pages will be automatically processed through Google Vision AI to perform the actual extraction. Figure 6 provides a visual representation of this process. Given that the analysed pages are particularly challenging, the results of this step will be then corrected by human experts, resulting in a high-quality OCR data set at the end of the pipeline.

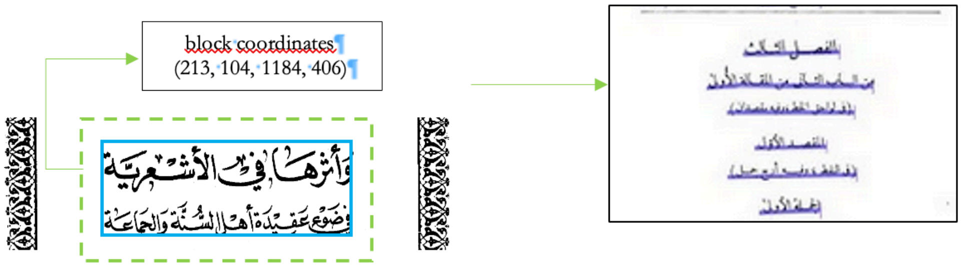

Figure 6 – Detail of a Google Vision AI output block segmentation and eScriptorium VRE / Kraken's line segmentation

A last important consideration must be made on the segmentation obtained by the Google Vision AI OCR which is a closed-source system and differs from the line segmentation of eScriptorium VRE / Kraken models (on the right in Figure 6). The goal of the block segmentation is to extract text regions with respect to the reading order; however, to train a model such as Kraken we need to segment regions (or blocks) into individual text lines. What is delineated is a two-step approach that, as demonstrated for example in the work of Martínek, Lenc, and Král [81], allows to determine logical text units and simplifies determining the reading order. Moreover, Google Vision AI block coordinates are a solid metadata basis to help in the block into line segmentation process that is made easier. In order to achieve this goal, once segmented the FPG with coordinates from Google Vision AI output could be aligned by employing the eScriptorium alignment functionalities or other text alignment tools such as Passim[18] already used thoroughly and with success in the KITAB project as well as extended in capabilities by another important project entitled Automatic Collation for Diversifying Corpora (ACDC) which also contributed the development of the Kraken OCR model for Arabic-script manuscripts recognition [111]. Once the alignment is completed a Kraken OCR model will be trained to segment and extract the text, that could be further post-processed.

Conclusion and Future Directions

In this study, we have presented and proposed our approach to help improve OCR capabilities for Arabic frontispieces in digital libraries. Recognising the unique challenges of Arabic script and the scarcity of specialised datasets tailored for the need of libraries, we aim to develop a comprehensive, high-quality dataset by processing approximately 140,000 historical documents. By employing advanced VLMs like Qwen-2vl-72B for page classification and OCR tools such as Google Vision AI for initial text extraction, we intend to create a reliable resource for training and benchmarking OCR algorithms. Our future work will focus on finalising this dataset, refining the OCR pipeline, and collaborating with linguistic experts to ensure accuracy. The following step will be to extend the evaluation of the text recognition performance of Google Vision AI and refine the results to achieve the best accuracy possible. A further step will test the implementation of different alignment solutions to connect the output generated with the segmentation structure requested by other open-source OCR, such as Kraken, to be trained properly.

In the multiliterate, multicultural, and multi-confessional European context, it could eventually contribute in reducing the social costs of religious illiteracy [92] by improving access to relevant

[18] Passim is the text reuse algorithm used by KITAB. The algorithm, through a naïve approach, identifies potential reuse by searching for shared text passages comparing text segments based on length and proximity thresholds (i.e., text exceeding a minimum word count and with a limited number of words between them). See https://kitab-project.org/methods/text-reuse and also https://github.com/dasmiq/passim.

sources to a wide array of users. The project could also contribute in (re-)discovering and expanding humanistic research areas, while attracting significant interest from the computer science world, which quickly achieves many of its challenges, often posed in an industrial framework, barely considering cultural, linguistic, historical, and religious nuances, as anticipated by Thaller already during the 1980s [118]. It should also be remembered that Humanities are able to pose cutting-edge questions to computer science (such as manuscript text recognition) that must then recalibrate and reconsider its approaches to address the complexity and richness of those datasets. This would stimulate both computer scientists and humanists to contribute in the development of an even more fertile research field for the ideation of innovative tool with multi-faceted values.

We believe this effort will significantly contribute to the training of OCR models for non-Latin script languages, such as Arabic, and to the preservation and accessibility of such texts in digital libraries, supporting advanced cataloguing and research initiatives. Moving forward, we plan to publicly release our curated frontispieces dataset, providing a valuable resource for the research community. With this dataset, we plan to train the open source Kraken OCR engine to develop a model specifically tailored for librarians and cataloguers usage, also aiming to improve OCR accuracy for these complex pages and to facilitate better metadata extraction and cataloguing. Additionally, by offering the dataset as a benchmark, we hope to support the evaluation and advancement of OCR systems focused on frontispiece recognition.

Lastly, in a more abstract sense we would like to suggest that the integration of this kind of multimodal AI models with library science and the librarians work could be considered as one attempt, in the new environment delineated by [101], to conciliate the architectural knowledge typical of the library with the oracular knowledge produced by new AI models.

References

[1] Adams, Richard M. 2022. ‘Defining Digital Pedagogy in Theological Libraries’. In *Digital Humanities and Libraries and Archives in Religious Studies: An Introduction*, edited by Clifford Anderson. De Gruyter. https://doi.org/10.1515/9783110536539-008.

[2] Adnan, Kiran, and Rehan Akbar. 2019. ‘Limitations of Information Extraction Methods and Techniques for Heterogeneous Unstructured Big Data’. *International Journal of Engineering Business Management* 11: 1–23. https://doi.org/10.1177/1847979019890771.

[3] Afli, Haithem, Loïc Barrault, and Holger Schwenk. 2016. ‘OCR Error Correction Using Statistical Machine Translation’. *Int. J. Comput. Linguistics Appl.* 7 (1): 175–91. https://loicbarrault.github.io/papers/afli_cicling2015.pdf.

[4] Ahmed, Muna, and Ali Abidi. 2019. ‘Review on Optical Character Recognition’. *International Research Journal of Engineering and Technology (IRJET)* 6 (6): 3666–69.

[5] AIB. Gruppo biblioteche digitali. 2005. ‘Manifesto per le biblioteche digitali’. https://www.aib.it/aib/cg/gbdigd05a.htm3.

[6] Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, et al. 2022. 'Flamingo: A Visual Language Model for Few-Shot Learning'. Paper presented at Advances in Neural Information Processing Systems. October 31. https://openreview.net/forum?id=EbMuimAbPbs.

[7] Al-Barhamtoshy, Hassanin M., Kamal M. Jambi, Mohsen A. Rashwan, and Sherif M. Abdou. 2023. 'An Arabic Manuscript Regions Detection, Recognition and Its Applications for OCRing'. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22 (1): 27:1-27:28. https://doi.org/10.1145/3532609.

[8] Alghyaline, Salah. 2022. 'A Printed Arabic Optical Character Recognition System Using Deep Learning'. *Journal of Computer Science* 18 (11): 1038–50. https://doi.org/10.3844/jcssp.2022.1038.1050.

[9] Al-Hajj, Ramy, Chafic Mokbel, and Laurence Likforman-Sulem. 2007. 'Combination of HMM-Based Classifiers for the Recognition of Arabic Handwritten Words'. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* 2: 959–63. https://api.semanticscholar.org/CorpusID:32687981.

[10] Al-Hashim, Amin G., and Sabri A. Mahmoud. 2010. 'Printed Arabic Text Database (PATDB) for Research and Benchmarking'. *Proceedings of the 9th WSEAS International Conference on Applications of Computer Engineering* (Stevens Point, Wisconsin, USA), ACE'10, 62–68. https://api.semanticscholar.org/CorpusID:2405250.

[11] Al-Kurdī, Muhammad Ṭāhir. 1939. *Tārīḫ al-ḫaṭṭ al-ʿarabī wa 'ādābi-hi*. 1st ed. Maktaba al-hilāl.

[12] Al-Ma'adeed, S., D. Elliman, and C.A. Higgins. 2004. 'A Data Base for Arabic Handwritten Text Recognition Research'. *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, 485–89. https://doi.org/10.1109/IWFHR.2002.1030957.

[13] Alma'adeed, S., C. Higgens, and D. Elliman. 2002. 'Recognition of Off-Line Handwritten Arabic Words Using Hidden Markov Model Approach'. *2002 International Conference on Pattern Recognition* 3: 481–84 vol.3. https://doi.org/10.1109/ICPR.2002.1047981.

[14] Al-Muhtaseb, H. A. 2010. 'Arabic Text Recognition of Printed Manuscripts : Efficient Recognition of off-Line Printed Arabic Text Using Hidden Markov Models, Bigram Statistical Language Model, and Post-Processing'. Ph.D., University of Bradford. http://hdl.handle.net/10454/4426.

[15] Alsheikh, Idris, and Masnizah Mohd. 2019. 'A Quranic Dataset for Text Recognition'. Paper presented at INCITEST 2019, Bandung, Indonesia.

Proceedings of the 1st International Conference on Informatics, Engineering, Science and Technology. https://doi.org/10.4108/eai.18-7-2019.2287842.

[16] Alūsī, ʿĀdil al-. 2008. *al-ḫaṭṭ al-ʿarabī: naša’tuhu wa taṭwwuruhu*. Maktaba al-dār al-ʿarabiyya li-l-kitāb.

[17] Amara, Najoua Essoukri Ben, Omar Mazhoud, Noura Bouzrara, and Noureddine Ellouze. 2005. 'ARABASE: A Relational Database for Arabic OCR Systems.' *Int. Arab J. Inf. Technol.* 2 (January): 259–66. https://api.semanticscholar.org/CorpusID:15268265.

[18] Anderson, Clifford, ed. 2022. 'Introduction'. In *Digital Humanities and Libraries and Archives in Religious Studies: An Introduction*. De Gruyter. https://doi.org/10.1515/9783110536539.

[19] Arani, Seyed, Ehsanollah Kabir, and Reza Ebrahimpour. 2019. 'Handwritten Farsi Word Recognition Using NN-Based Fusion of HMM Classifiers with Different Types of Features'. *International Journal of Image and Graphics* 19 (January): 1950001. https://doi.org/10.1142/S0219467819500013.

[20] Badry, Mahmoud, Hesham Hassan, Hanaa Bayomi, and Hussien Oakasha. 2018. 'QTID: Quran Text Image Dataset'. *International Journal of Advanced Computer Science and Applications (IJACSA)* 9 (3): 3. https://doi.org/10.14569/IJACSA.2018.090351.

[21] Baghshah, M. Soleymani, S. Bagheri Shouraki, and S. Kasaei. 2005. 'A Novel Fuzzy Approach to Recognition of Online Persian Handwriting'. September 1, 268–73. https://doi.org/10.1109/ISDA.2005.13.

[22] Bai, Jinze, Shuai Bai, Shusheng Yang, et al. 2023. 'Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond'. arXiv:2308.12966. Preprint, arXiv, October 13. https://doi.org/10.48550/arXiv.2308.12966.

[23] Balāḏurī, ʾAḥmad bin Yaḥyà al-. 1987. *Futūḥ al-buldān*. Mu’assasat al-maʿārif.

[24] Bassil, Youssef, and Mohammad Alwani. 2012. *OCR Post-Processing Error Correction Algorithm Using Google Online Spelling Suggestion*.

[25] Bataineh, Bilal. 2017. 'A Printed PAW Image Database of Arabic Language for Document Analysis and Recognition'. *Journal of ICT Research and Applications* 11 (August): 199–211. https://doi.org/10.5614/itbj.ict.res.appl.2017.11.2.6.

[26] Bergamaschi, Sonia, Stefania De Nardis, Riccardo Martoglia, et al. 2022. 'Novel Perspectives for the Management of Multilingual and Multialphabetic Heritages through Automatic Knowledge Extraction: The DigitalMaktaba Approach'. *Sensors* 22 (11): 11. https://doi.org/10.3390/s22113995.

[27] Bergamaschi, Sonia, Riccardo Martoglia, Federico Ruozzi, et al. 2021. 'Preserving and Conserving Culture: First Steps towards a Knowledge Extractor and

Cataloguer for Multilingual and Multi-Alphabetic Heritages'. *Proceedings of the Conference on Information Technology for Social Good* (New York, NY, USA), GoodIT '21, September 9, 301–4. https://doi.org/10.1145/3462203.3475927.

[28] Bian, Xiaohang, Bo Qin, Xiaozhe Xin, Jianwu Li, Xuefeng Su, and Yanfeng Wang. 2022. 'Handwritten Mathematical Expression Recognition via Attention Aggregation Based Bi-Directional Mutual Learning'. arXiv:2112.03603. Preprint, arXiv, February 23. https://doi.org/10.48550/arXiv.2112.03603.

[29] Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. 2016. *Digital_Humanities*. MIT Press.

[30] Bushofa, Bmf, and M Spann. 1995. 'Segmentation and Recognition of Printed Arabic Characters.' *Procedings of the British Machine Vision Conference 1995*, 543–52. https://doi.org/10.5244/C.9.54.

[31] Buzzetti, Dino. 2006. 'Biblioteche digitali e oggetti digitali complessi: esaustività e funzionalità nella conservazione'. *Archivi informatici per il patrimonio culturale*, Contributi del centro linceo interdisciplinare 'Beniamino Segre', vol. 114: 36.

[32] Buzzetti, Dino. 2023. 'Towards an Operational Approach to Computational Text Analysis'. In *On Making in the Digital Humanities: The Scholarship of Digital Humanities Development in Honour of John Bradley*, First, edited by Julianne Nyhan, Geoffrey Rockwell, Stéfan Sinclair, and Alexandra Ortolja-Baird. UCL Press.
https://press.uchicago.edu/ucp/books/book/distributed/O/bo208645506.html.

[33] Chabchoub, Fatma, Yousri Kessentini, Slim Kanoun, and Véronique Eglin. 2016. 'SmartATID: A Mobile Captured Arabic Text Images Dataset for Multi-Purpose Recognition Tasks'. Paper presented at Internation Conference in Frontiers on Handwriting Recognition, Shenzhen, China. *Internation Conference in Frontiers on Handwriting Recognition*, Internation Conference in Frontiers on Handwriting Recognition, October. https://hal.archives-ouvertes.fr/hal-01403764.

[34] Chen, Zhe, Jiannan Wu, Wenhai Wang, et al. 2024. 'InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks'. arXiv:2312.14238. Preprint, arXiv, January 15. https://doi.org/10.48550/arXiv.2312.14238.

[35] Chiron, Guillaume, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. 'Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information'. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–4. https://doi.org/10.1109/JCDL.2017.7991582.

[36] Chtourou, Imen, Ahmed Cheikh Rouhou, Faten Kallel Jaiem, and Slim Kanoun. 2015. 'ALTID : Arabic/Latin Text Images Database for Recognition Research'.

2015 13th International Conference on Document Analysis and Recognition (ICDAR), August 1, 836–40. https://doi.org/10.1109/ICDAR.2015.7333879.

[37] Davidson, R., and R. Hopely. 1997. 'Arabic and Persian OCR Training and Test Data Sets'. *Proc. of Symp. on Document Image Understanding Technology*, 303–7.

[38] Dehghan, M., and K. Faez. 1997. 'Farsi Handwritten Character Recognition with Moment Invariants'. *Proceedings of 13th International Conference on Digital Signal Processing* 2 (July): 507–10 vol.2. https://doi.org/10.1109/ICDSP.1997.628387.

[39] Dehghani, A., F. Shabani, and P. Nava. 2001. 'Off-Line Recognition of Isolated Persian Handwritten Characters Using Multiple Hidden Markov Models'. April 1, 0506–0506. https://doi.org/10.1109/ITCC.2001.918847.

[40] Ding, Yihao, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. 'PDF-VQA: A New Dataset for Real-World VQA on PDF Documents'. *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part VI* (Berlin, Heidelberg), September 18, 585–601. https://doi.org/10.1007/978-3-031-43427-3_35.

[41] Djeddi, Chawki, Abdeljalil Gattal, Labiba Souici-Meslati, Imran Siddiqi, Youcef Chibani, and Haikal El Abed. 2014. 'LAMIS-MSHD: A Multi-Script Offline Handwriting Database'. *2014 14th International Conference on Frontiers in Handwriting Recognition*, September, 93–97. https://doi.org/10.1109/ICFHR.2014.23.

[42] Dreuw, Philippe, Stephan Jonas, and Hermann Ney. 2008. 'White-Space Models for Offline Arabic Handwriting Recognition'. *2008 19th International Conference on Pattern Recognition*, December, 1–4. https://doi.org/10.1109/ICPR.2008.4761841.

[43] Durand, Olivier, Angela Daiana Langone, and Giuliano Mion. 2010. *Corso di arabo contemporaneo*. Hoepli.

[44] Elleuch, Mohamed, Rania Maalej, and Monji Kherallah. 2016. 'A New Design Based-SVM of the CNN Classifier Architecture with Dropout for Offline Arabic Handwritten Recognition'. *Procedia Computer Science*, International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA, vol. 80 (January): 1712–23. https://doi.org/10.1016/j.procs.2016.05.512.

[45] Elzobi, Moftah, Ayoub Al-Hamadi, Zaher Al Aghbari, and Laslo Dings. 2013. 'IESK-ArDB: A Database for Handwritten Arabic and an Optimized Topological Segmentation Approach'. *International Journal on Document Analysis and Recognition (IJDAR)* 16 (3): 295–308. https://doi.org/10.1007/s10032-012-0190-z.

[46] Escolano Rodríguez, Elena, Adelaida Caro Martín, Judit Fejes, et al. 2022. *ISBD International Standard Bibliographic Description : 2021 Update to the 2011 Consolidated Edition*. February. https://repository.ifla.org/handle/20.500.14598/1939.

[47] Experimental HumanitiesLabat theIliffSchoolof Theology. 2022. 'Library as Interface for Digital Humanities'. In *Digital Humanities and Libraries and Archives in Religious Studies: An Introduction*, edited by Clifford B. Anderson. De Gruyter. https://doi.org/10.1515/9783110536539-010.

[48] Faḍā'ilī, Ḥabīb Allah. 2002. *'aṭlas al-ḫaṭṭ wa al-ḫuṭūṭ*. 2nd ed. Maktaba al-ṭalās.

[49] Faizullah, Safiullah, Muhammad Sohaib Ayub, Sajid Hussain, and Muhammad Asad Khan. 2023. 'A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges'. *Applied Sciences* 13 (7): 7. https://doi.org/10.3390/app13074584.

[50] Farooq, Faisal, Venu Govindaraju, and Michael Perrone. 2005. 'Pre-Processing Methods for Handwritten Arabic Documents'. August 1, 267–71. https://doi.org/10.1109/ICDAR.2005.191.

[51] Fasha, Mohammad, Bassam Hammo, Nadim Obeid, and Jabir Widian. 2020. 'A Hybrid Deep Learning Model for Arabic Text Recognition'. *(IJACSA) International Journal of Advanced Computer Science and Applications* 11 (8): 122–30. https://doi.org/10.14569/issn.2156-5570.

[52] Garbini, Giovanni, and Olivier Durand. 1994. *Introduzione alle lingue semitiche*. Studi sul Vicino Oriente antico. Paideia.

[53] Ġarīd al-Šayḫ, Muḥammad. 2010. *al-Muʿğam fī al-luġa wa al-naḥw wa al-ṣarf wa al-'iʿrāb wa al-muṣṭalaḥāt al-ʿilmiyya wa al-falsafiyya wa al-qānūniyya wa al-ḥadīṯa*. Vol. 5. Mu'assasa al-nuḫba li-l-ta'līf wa al-tarğama wa al-našr.

[54] Gemini Team. 2024. 'Gemini: A Family of Highly Capable Multimodal Models'. arXiv:2312.11805. Preprint, arXiv, April 2. https://doi.org/10.48550/arXiv.2312.11805.

[55] George, Alain. 2010. *The Rise of Islamic Calligraphy*. Saqi Books.

[56] Graves, Alex. 2012. 'Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural Networks'. In *Guide to OCR for Arabic Scripts*, edited by Volker Märgner and Haikal El Abed. Springer. https://doi.org/10.1007/978-1-4471-4072-6_12.

[57] Gruppo di lavoro sulle biblioteche digitali (GBDIG) and Gruppo di studio sulle tecnologie dell'informazione nelle biblioteche e biblioteche digitali. 2020. 'Nuovo

Manifesto per le biblioteche digitali'. AIB WEB. https://www.aib.it/documenti/nuovo-manifesto-per-le-biblioteche-digitali/.

[58] Guerrini, Mauro. 2005. *Verso nuovi principi e nuovi codici di catalogazione*. Edited by Carlo Bianchini. With Carlo Bianchini, Pino Buizza, Carlo Ghilli, Antonella Novelli, Lucia Sardo, and Alberto Petrucciani. Studi Bibliografici. Edizioni Sylvestre Bonnard.

[59] Hamdani, Mahdi, Patrick Doetsch, Michal Kozielski, Amr Mousa, and Hermann Ney. 2014. 'The RWTH Large Vocabulary Arabic Handwriting Recognition System'. Paper presented at Proceedings - 11th IAPR International Workshop on Document Analysis Systems, DAS 2014. April 30. https://doi.org/10.1109/DAS.2014.61.

[60] Handelman, Matthew. 2022. 'A Messianic Theory of Digital Knowledge: On Positivism and Visualizing Rosenzweig's Archive'. In *Digital Humanities and Libraries and Archives in Religious Studies: An Introduction*, edited by Clifford B. Anderson. De Gruyter. https://doi.org/10.1515/9783110536539-004.

[61] Hegghammer, Thomas. 2022. 'OCR with Tesseract, Amazon Textract, and Google Document AI: A Benchmarking Experiment'. *Journal of Computational Social Science* 5 (1): 861–82. https://doi.org/10.1007/s42001-021-00149-1.

[62] Ibn al-Nadīm, Muḥammad bin ʾIsḥāq Abū al-Farağ bin ʾAbī Yaʿqūb al-Nadīm. 1997. *al-Fihrist*. Edited by Ibrāhīm Ramaḍān. Dār al-maʿrifa.

[63] Ibn Durayd, Abu Bakr Muḥmmad bin al-Ḥasan. 1991. *al-ištiqāq*. Edited by ʿAbd al-Salām Muḥammad Hārūn. Dār al-Ǧīl.

[64] Ibn ʿAbd Rabbih, Abu ʿUmar Aḥmad. 1983. *al-ʿiqd al-farīd*. Edited by A.M. Tarḥīnī. Vol. 4. Dar al-kutub al-ʿilmiyya.

[65] Jaiem, Faten Kallel, Slim Kanoun, Maher Khemakhem, Haikal El Abed, and Jihain Kardoun. 2013. 'Database for Arabic Printed Text Recognition Research'. In *Image Analysis and Processing – ICIAP 2013*, edited by Alfredo Petrosino. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-642-41181-6_26.

[66] Kazemi, Maziar, Muhammad Yousefnezhad, and Saber Nourian. 2015. 'A New Approach in Persian Handwritten Letters Recognition Using Error Correcting Output Coding'. *Journal of Advances in Computer Research* 6 (4): 107–24. DOI: https://doi.org/10.48550/arXiv.1604.07554.

[67] Kew, Jhonathan. 2005. *Notes on Some Unicode Arabic Characters: Recommendations for Usage*.

[68] Khalaf, Khatatneh, Ibrahiem Emary, and Basem Rifai. 2006. 'Probabilistic Artificial Neural Network For Recognizing the Arabic Hand Written Characters'.

Journal of Computer Science 2 (12): 879–84. https://doi.org/10.3844/jcssp.2006.879.884.

[69] Khorsheed, M.S., and W.F. Clocksin. 2000. 'Multi-Font Arabic Word Recognition Using Spectral Features'. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000* 4: 543–46. https://doi.org/10.1109/ICPR.2000.902977.

[70] Khosrobeygi, Z., H. Veisi, H. R. Ahmadi, and H. Shabanian. 2020. 'A Rule-Based Post-Processing Approach to Improve Persian OCR Performance'. *Scientia Iranica* 27 (6): 3019–33. https://doi.org/10.24200/sci.2020.53435.3267.

[71] Khosrobeygi, Zohreh, Hadi Veisi, Ehsan Hoseinzade, and Hanieh Shabanian. 2022. 'Persian Optical Character Recognition Using Deep Bidirectional Long Short-Term Memory'. *Applied Sciences* 12 (22): 22. https://doi.org/10.3390/app122211760.

[72] Kiessling, Benjamin, Gennady Kurin, Matthew Thomas Miller, and Kader Smail. 2021. 'Advances and Limitations in Open Source Arabic-Script OCR: A Case Study'. *Digital Studies / Le Champ Numérique* 11 (1). https://doi.org/10.16995/dscn.8094.

[73] Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. 'BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation'. arXiv:2201.12086. Preprint, arXiv, February 15. https://doi.org/10.48550/arXiv.2201.12086.

[74] Li, Zongxia, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. 'A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges'. arXiv:2501.02189. Preprint, arXiv, April 6. https://doi.org/10.48550/arXiv.2501.02189.

[75] Liu, Haotian, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. 'Improved Baselines with Visual Instruction Tuning'. arXiv:2310.03744. Preprint, arXiv, May 15. https://doi.org/10.48550/arXiv.2310.03744.

[76] Liu, Xiaojing, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. 'Graph Convolution for Multimodal Information Extraction from Visually Rich Documents'. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, edited by Anastassia Loukina, Michelle Morales, and Rohit Kumar. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-2005.

[77] Lorigo, L., and Venu Govindaraju. 2005. 'Segmentation and Pre-Recognition of Arabic Handwriting'. 2005 (January): 605-609 Vol. 2. https://doi.org/10.1109/ICDAR.2005.207.

[78] Luqman, Hamzah, Sabri A. Mahmoud, and Sameh Awaida. 2014. 'KAFD Arabic Font Database'. *Pattern Recognition* 47 (6): 2231–40. https://doi.org/10.1016/j.patcog.2013.12.012.

[79] Maalej, Rania, and Monji Kherallah. 2018. 'Convolutional Neural Network and BLSTM for Offline Arabic Handwriting Recognition'. *2018 International Arab Conference on Information Technology (ACIT)*, November, 1–6. https://doi.org/10.1109/ACIT.2018.8672667.

[80] Maalej, Rania, and Monji Kherallah. 2019. 'Maxout into MDLSTM for Offline Arabic Handwriting Recognition'. *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III* (Berlin, Heidelberg), December 12, 534–45. https://doi.org/10.1007/978-3-030-36718-3_45.

[81] Martínek, Jiří, Ladislav Lenc, and Pavel Král. 2020. 'Building an Efficient OCR System for Historical Documents with Little Training Data'. *Neural Computing and Applications* 32 (23): 17209–27. https://doi.org/10.1007/s00521-020-04910-x.

[82] Martoglia, Riccardo, Sonia Bergamaschi, Federico Ruozzi, Matteo Vanzini, Luca Sala, and Riccardo Amerigo Vigliermo. 2023. 'Knowledge Extraction, Management and Long-Term Preservation of Non-Latin Cultural Heritages - Digital Maktaba Project Presentation'. In *Proceedings of the 19th Conference on Information and Research Science Connecting to Digital and Library Science*, edited by Bardi Alessia, Falcon Alex, Ferilli Stefano, Marchesin Stefano, and Redavid Domenico, vol. 3365. CEUR Workshop Proceedings. CEUR. https://ceur-ws.org/Vol-3365/#short11.

[83] Martoglia, Riccardo, Luca Sala, Matteo Vanzini, and Riccardo Amerigo Vigliermo. 2022a. 'A Tool for Semiautomatic Cataloguing of an Islamic Digital Library: A Use Case from the Digital Maktaba Project'. In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022), Berlin, Germany, Sept. 19th-23rd, 2022*, edited by Adrian Paschke, Georg Rehm, Clemens Neudecker, and Lydia Pintscher, vol. 3234. CEUR Workshop Proceedings. CEUR-WS.org. https://ceur-ws.org/Vol-3234/paper1.pdf.

[84] Moghaddam, Reza Farrahi, Mohamed Cheriet, Mathias M. Adankon, Kostyantyn Filonenko, and Robert Wisnovsky. 2010. 'IBN SINA: A Database for Research on Processing and Understanding of Arabic Manuscripts Images'. *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems - DAS'10*, 11–18. https://doi.org/10.48550/arXiv.1604.07554.

[85] Mohd, Masnizah, Faizan Qamar, Idris Al-Sheikh, and Ramzi Salah. 2021. 'Quranic Optical Text Recognition Using Deep Learning Models'. *IEEE Access* 9: 38318–30. https://doi.org/10.1109/ACCESS.2021.3064019.

[86] Morriello, Rossana. 2020. 'Birth and Development of Data Librarianship'. *Jlis.it* 11: 1–15. https://doi.org/10.4403/jlis.it-12653.

[87] Mostafa, Aly, Omar Mohamed, Ali Ashraf, et al. 2022. 'An End-to-End OCR Framework for Robust Arabic-Handwriting Recognition Using a Novel Transformers-Based Model and an Innovative 270 Million-Words Multi-Font Corpus of Classical Arabic with Diacritics'. arXiv:2208.11484. Preprint, arXiv, August 26. http://arxiv.org/abs/2208.11484.

[88] Mowlaei, A., K. Faez, and A.T. Haghighat. 2002. 'Feature Extraction with Wavelet Transform for Recognition of Isolated Handwritten Farsi/Arabic Characters and Numerals'. *2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No.02TH8628)* 2: 923–26. https://doi.org/10.1109/ICDSP.2002.1028240.

[89] Mozaffari, S., K. Faez, and H.R. Kanan. 2004. 'Recognition of Isolated Handwritten Farsi/Arabic Alphanumeric Using Fractal Codes'. *6th IEEE Southwest Symposium on Image Analysis and Interpretation, 2004.*, March, 104–8. https://doi.org/10.1109/IAI.2004.1300954.

[90] Mozaffari, Saeed, Haikal El Abed, Volker Märgner, Karim Faez, and Ali Amirshahi. 2008. 'IfN/Farsi-Database: A Database of Farsi Handwritten City Names'. *International Conference on Frontiers in Handwriting Recognition*.

[91] Mustapha, Ismail B., Shafaatunnur Hasan, Hatem Nabus, and Siti Mariyam Shamsuddin. 2022. 'Conditional Deep Convolutional Generative Adversarial Networks for Isolated Handwritten Arabic Character Generation'. *Arabian Journal for Science and Engineering* 47 (2): 1309–20. https://doi.org/10.1007/s13369-021-05796-0.

[92] Naso, Paolo. 2014. 'Rapporto sull'analfabetismo religioso in Italia'. In *I costi sociali dell'analfabetismo religioso*, edited by Alberto Melloni. Il Mulino.

[93] Naz, Saeeda, Arif Iqbal Umar, Syed H. Shirazi, Saeed B. Ahmed, Muhammad I. Razzak, and Imran Siddiqi. 2015. 'Segmentation Techniques for Recognition of Arabic-like Scripts: A Comprehensive Survey'. *Education and Information Technologies*, Springer Journal of Education and Information Technologies, vol. 21 (5): 1–20. https://doi.org/10.1007/s10639-015-9377-5.

[94] Nguyen, Kim-Anh Laura. 2024. 'Document Understanding with Deep Learning Techniques. Document and Text Processing.' Sorbonne Université, 2024.

https://theses.hal.science/tel-04626992/file/140733_NGUYEN_2024_archivage.pdf.

[95] Nguyen, Thi Tuyet Hai, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. ‘Survey of Post-OCR Processing Approaches’. *ACM Computing Surveys* 54 (6): 1–37. https://doi.org/10.1145/3453476.

[96] OpenAI, Josh Achiam, Steven Adler, et al. 2024. ‘GPT-4 Technical Report’. arXiv:2303.08774. Preprint, arXiv, March 4. https://doi.org/10.48550/arXiv.2303.08774.

[97] Pechwitz, Mario, Samia Snoussi Maddouri, Volker Märgner, Noureddine Ellouze, and Hamid Amiri. 2002. ‘IFN/ENIT - Database of Handwritten Arabic Words’. *In Proc. of CIFED 2002*, 129–36.

[98] Qalqašandī, Abu al-ʿAbbās ’Aḥmad al-. 1913. *Ṣubḥ al-’aʿšà*. Vol. 3. Al-Maṭbaʿa al-’amīriyya.

[99] Radford, Alec, Jong Wook Kim, Chris Hallacy, et al. 2021. ‘Learning Transferable Visual Models From Natural Language Supervision’. arXiv:2103.00020. Preprint, arXiv, February 26. https://doi.org/10.48550/arXiv.2103.00020.

[100] Ramdan, Jabril, Khairuddin Omar, Mohammad Faidzul, and Ali Mady. 2013. ‘Arabic Handwriting Data Base for Text Recognition’. *Procedia Technology* 11: 580–84. https://doi.org/10.1016/j.protcy.2013.12.231.

[101] Roncaglia, Gino. 2023. *L’architetto e l’oracolo: Forme Digitali Del Sapere Da Wikipedia a ChatGPT*. Editori Laterza.

[102] Russell, John E., and Merinda Kaye Hensley. 2017. ‘Beyond Buttonology: Digital Humanities, Digital Pedagogy, and the ACRL Framework’. *College & Research Libraries News* 78 (11): 11. https://doi.org/10.5860/crln.78.11.588.

[103] Saddami, Khairun, Khairul Munadi, and Fitri Arnia. 2015. ‘A Database of Printed Jawi Character Image’. *2015 Third International Conference on Image Information Processing (ICIIP)*, December, 56–59. https://doi.org/10.1109/ICIIP.2015.7414740.

[104] Sadri, Javad, Mohammad Reza Yeganehzad, and Javad Saghi. 2016. ‘A Novel Comprehensive Database for Offline Persian Handwriting Recognition’. *Pattern Recognition* 60 (C): 378–93. https://doi.org/10.1016/j.patcog.2016.03.024.

[105] Safabaksh, Reza, Ali Reza Ghanbarian, and Golnaz Ghiasi. 2013. ‘HaFT: A Handwritten Farsi Text Database’. *2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP)*, September, 89–94. https://doi.org/10.1109/IranianMVIP.2013.6779956.

[106] Safadi, Yasin Hamid. 1978. *Islamic Calligraphy*. Thames and Hudson.

[107] Saloum, Said S. 2021. 'DAD: A Detailed Arabic Dataset for Online Text Recognition and Writer Identification, a New Type'. *Journal of Computer Science* 17 (1): 19–32. https://doi.org/10.3844/jcssp.2021.19.32.

[108] Schimmel, Annemarie, and Barbara Rivolta. 1992. *Islamic Calligraphy*. Brill Archive.

[109] Schwartz, Christine. 2022. 'Using XQuery and XSLT to Build an Aggregation of Metadata Records for Religious Texts and Non-Print Items'. In *Digital Humanities and Libraries and Archives in Religious Studies: An Introduction*, edited by Clifford B. Anderson. De Gruyter. https://doi.org/10.1515/9783110536539-007.

[110] Slimane, Fouad, Rolf Ingold, Slim Kanoun, Adel Alimi, and Jean Hennebert. 2009. 'A New Arabic Printed Text Image Database and Evaluation Protocols'. *10th International Conference on Document Analysis and Recognition*, 946–50. https://doi.org/10.1109/ICDAR.2009.155.

[111] Smith, David A., Jacob Murel, Jonathan Parkes Allen, and Matthew Thomas Miller. 2023. 'Automatic Collation for Diversifying Corpora: Commonly Copied Texts as Distant Supervision for Handwritten Text Recognition'. In *Proceedings of the Computational Humanities Research Conference 2023*, edited by Artjoms Šeļa, Fotis Jannidis, and Iza Romanowska, vol. 3558. CEUR Workshop Proceedings. CEUR. https://ceur-ws.org/Vol-3558/#paper1708.

[112] Smith, David, and Ryan Cordell. 2018. 'A Research Agenda for Historical and Multilingual Optical Character Recognition - DRS'. https://repository.library.northeastern.edu/files/neu:f1881m035.

[113] Stokes, Peter, Benjamin Kiessling, Daniel Stökl Ben Ezra, R. Tissot, and E.H. Gargem. 2021. 'The EScriptorium VRE for Manuscript Cultures'. *Classics@ Journal, Ancient Manuscripts and Virtual Research Environments* 18.

[114] Sturgeon, Donald. 2021. 'Chinese Text Project: A Dynamic Digital Library of Premodern Chinese'. *Digital Scholarship in the Humanities* 36 (Supplement_1): 101–12. https://doi.org/10.1093/llc/fqz046.

[115] Ṣubḥī Murād, Ḥassān. 2003. *Tārīḫ al-ḫaṭṭ al-ʿarabī: bayna al-māḍī wa-l-ḥāḍir*. Al-dār al-ǧamāhīriyya li-l-na wa al-tawzīʿwa al-'iʿlān.

[116] Sullutrone, Giovanni, Riccardo Amerigo Vigliermo, Luca Sala, and Sonia Bergamaschi. 2024. 'Sensitive Topics Retrieval in Digital Libraries: A Case Study of Ḥadīṯ Collections'. In *Linking Theory and Practice of Digital Libraries*, edited by Apostolos Antonacopoulos, Annika Hinze, Benjamin Piwowarski, et al. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72440-4_5.

[117] Tang, Jingqun, Chunhui Lin, Zhen Zhao, et al. 2024. 'TextSquare: Scaling up Text-Centric Visual Instruction Tuning'. arXiv:2404.12803. Preprint, arXiv, April 19. https://doi.org/10.48550/arXiv.2404.12803.

[118] Thaller, Manfred. 1989. 'The Need for a Theory of Historical Computing'. In *History and Computing II*, edited by P. Denley, S. Folgevik, and C. Harvey, vol. 2. Manchester Unversity Press.

[119] Timsari, Bijan, and Hamid Fahimi. 1996. *Morphological Approach to Character Recognition in Machine-Printed Persian Words*. 2660 (March): 184–91. https://doi.org/10.1117/12.234724.

[120] Torki, Marwan, Mohamed Husseiny, Ahmed Elsallamy, Mahmoud Fayyaz, and Shehab Yaser. 2014. *Window-Based Descriptors for Arabic Handwritten Alphabet Recognition: A Comparative Study on a Novel Dataset*. https://doi.org/10.48550/arXiv.1411.3519.

[121] Touj, Sameh Masmoudi, Najoua Essoukri Ben Amara, and Hamid Amiri. 2007. 'A Hybrid Approach for Off-Line Arabic Handwriting Recognition Based on a Planar Hidden Markov Modeling'. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* 2: 964–68. https://dl.acm.org/doi/10.5555/1304596.1304932.

[122] Vigliermo, Riccardo Amerigo, Giovanni Sullutrone, Sonia Bergamaschi, and Luca Sala. 2025. 'Proposing a Comprehensive Dataset for Arabic Script OCR in the Context of Digital Libraries and Religious Archives (Extended Abstract)'. In *Proceedings of the 21st Conference on Information and Research Science Connecting to Digital and Library Science*, edited by Marcella Cornia, Giorgio Maria Di Nunzio, Donatella Firmani, et al., vol. 3937. CEUR Workshop Proceedings. CEUR. https://ceur-ws.org/Vol-3937/#short5.

[123] Wang, Hao, Qingxuan Wang, Yue Li, Changqing Wang, Chenhui Chu, and Rui Wang. 2023. 'DocTrack: A Visually-Rich Document Dataset Really Aligned with Human Eye Movement for Machine Reading'. arXiv:2310.14802. Preprint, arXiv, October 23. https://doi.org/10.48550/arXiv.2310.14802.

[124] Wang, Peng, Shuai Bai, Sinan Tan, et al. 2024. 'Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution'. arXiv:2409.12191. Preprint, arXiv, October 3. https://doi.org/10.48550/arXiv.2409.12191.

[125] Xu, Derong, Wei Chen, Wenjun Peng, et al. 2023. 'Large Language Models for Generative Information Extraction: A Survey'. arXiv:2312.17617. Preprint, arXiv, December 29. https://doi.org/10.48550/arXiv.2312.17617.

[126] Yousfi, Sonia, Sid-Ahmed Berrani, and Christophe Garcia. 2015. 'ALIF: A Dataset for Arabic Embedded Text Recognition in TV Broadcast'. August 1, 1221–25. https://doi.org/10.1109/ICDAR.2015.7333958.

[127] Zahedi, Morteza, and Saeideh Eslami. 2011. 'Farsi/Arabic Optical Font Recognition Using SIFT Features'. *Procedia Computer Science*, World Conference on Information Technology, vol. 3 (January): 1055–59. https://doi.org/10.1016/j.procs.2010.12.173.

[128] Zayene, Oussama, Jean Hennebert, Sameh Masmoudi Touj, Rolf Ingold, and Najoua Essoukri Ben Amara. 2015. 'A Dataset for Arabic Text Detection, Tracking and Recognition in News Videos- AcTiV'. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, August, 996–1000. https://doi.org/10.1109/ICDAR.2015.7333911.

[129] Zayene, Oussama, Sameh Masmoudi Touj, Jean Hennebert, Rolf Ingold, and Najoua Essoukri Ben Amara. 2018. 'Open Datasets and Tools for Arabic Text Detection and Recognition in News Video Frames'. *Journal of Imaging* 4 (2): 32. https://doi.org/10.3390/jimaging4020032.

[130] ʿAfīfī, F.S. 1980. *Našāt wa taṭawwur al-kitāba al-ḫaṭṭiyya al-ʿarabiyya wa dawri-ha al-ṯaqāfī wa al-iǧtimāʿī*. Wikāla al-maṭbūʿāt.