The Codes Talk: Automated Extraction and Normalization of ISBNs for Metadata Integration

Sania Aftar[1,*,†], Riccardo Amerigo Vigliermo[2] and Sonia Bergamaschi[1]

[1]University of Modena and Reggio Emilia
[2]Mem s.r.l., via Cesare Boldrini 24, Bologna (BO), 40121

Abstract

Digitizing library collections is vital for preserving cultural heritage, yet Arabic and bilingual texts often suffer from incomplete metadata particularly ISBNs, hindering catalog integration. OCR noise, mixed numeral systems (Arabic-Indic, Persian, Western) and bidirectional text complexities further challenge reliable extraction from Arabic script materials. This paper introduces a lightweight, fully automated pipeline for extracting, normalizing and matching ISBNs from unstructured Arabic bibliographic metadata. The pipeline combines regex based extraction with specialized normalization: numeral system conversion, character level error correction via confusion dictionary, bidirectionality resolution and checksum validation, with fallback matching on titles and authors when ISBNs are missing. Pipeline is evaluated on a large collection of Arabic books, which substantially outperformed regex only baseline and demonstrated high precision and recall. Catalog reconciliation revealed minimal overlap with existing holdings while identifying a substantial proportion of new materials with valid ISBNs, underscoring the systematic under representation of Arabic script materials in Western library collections. This scalable, cost free and reproducible approach proves particularly valuable for resource constrained institutions managing script diverse collections, advancing metadata standardization and cultural heritage preservation.

Keywords

Digital Libraries, ISBN Normalization, Arabic Scripts, Metadata Integration, OCR Noise, Automated Extraction

1. Introduction

The rapid growth of digital libraries has transformed how knowledge is preserved, organized, and accessed. Structured bibliographic metadata is central to this transformation, enabling efficient cataloging, discovery, and integration across repositories. Among these metadata elements, the International Standard Book Number (ISBN) plays a critical role as a unique and globally recognized identifier for books. Ensuring accurate and consistent ISBN records is essential for supporting interoperability, reducing duplication, and enhancing resource discoverability in large scale and multilingual digital collections [1]. However, such repositories often suffer from incomplete, inconsistent, or erroneous identifiers, especially when metadata originates from legacy systems, low quality MARC records or OCR processed files [2, 3]. While automated extraction and normalization methods can achieve high levels of accuracy and scalability [4], challenges remain when dealing with complex scripts, typographic noise, and mixed numeral systems issues, particularly pronounced in collections containing Arabic script documents. In this specific case, large religious digital archives often include texts from various periods, many predating the introduction of ISBNs, which are a relatively recent and localized development[1].

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.
†These authors contributed equally.
327178@studenti.unimore.it (S. Aftar); riccardo.vigliermo@mem-dh.it (R. A. Vigliermo); sonia.bergamaschi@unimore.it (S. Bergamaschi)
https://unimore.unifind.cineca.it/resource/person/260520 (S. Aftar); https://www.linkedin.com/in/riccardo-amerigo-vigliermo-581645164/details/publications/;https://www.mem-dh.it/ (R. A. Vigliermo); https://www.unimore.it/it/ateneo/professori-emeriti/profssa-sonia-bergamaschi (S. Bergamaschi)
0000-0001-8151-8941 (S. Aftar); 0000-0001-9914-3295 (R. A. Vigliermo); 0000-0001-8087-6587 (S. Bergamaschi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]The ISBN system was developed in 1965-1967 for UK bookseller W. H. Smith's computerized warehouse. It consists of a unique 13-digit identifier (formerly 10) structured to identify the publisher, title, edition, and format, assigned to books for

CEUR-WS.org/Vol-4240/short11.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-08-20

From a Library Science standpoint, it is also worth reminding that the IFLA's International Standard of Bibliographic Description (ISBD) guidelines designate ISBN as "mandatory if available" (coded MA). Thus, there is no requirement for publications to have an ISBN [5]. These factors undermine automated catalog integration and metadata reconciliation, highlighting the need for language-sensitive robust solutions. To address these challenges, we propose a fully automated pipeline for extracting, normalizing, and matching ISBN-based metadata from unstructured multilingual historical book records. The pipeline combines rule-based normalization, numeral system conversion, structural validation, and character-level OCR error correction to handle diverse ISBN notations, including Arabic-Indic and Persian digits. Regex-based parsing with heuristics extracts candidates from unstructured metadata, while fallback matching on titles and authors addresses missing or malformed ISBNs. The system is designed for scalable integration with external catalogs.

2. Literature Review

Metadata quality in digital libraries has been extensively studied, with common issues like inconsistency, incompleteness, and errors affecting discoverability and catalog integration. Projects such as the University of Houston's metadata audit [6], ETD remediation efforts [7], and broader initiatives like Europeana and DPLA have employed manual, rule based, and semi automated strategies [8, 9], while recent work explores Machine Learning and AI for automating metadata error detection and enrichment [10, 11].

However, OCR implementation introduced new challenges in metadata extraction for complex scripts like Arabic, including script handling, segmentation and recognition issues and bidirectionality problems with Arabic-Indic digits [12, 13]. Open-source OCR tools show both advancement and limitations when applied to historical and degraded Arabic script texts [14], with OCR errors hindering reliable extraction of resource identifiers like ISBNs and affecting catalog integration.

To our knowledge, no prior work jointly addresses OCR correction, multilingual metadata normalization, and validation in noisy historical digital library records. This study fills that gap by proposing an automated, language sensitive pipeline.

3. Methodology

The primary objective of this initiative was to determine which books from a large donated collection were already present in the La Pira Library's online catalog and which represented new materials requiring cataloging. To achieve this, we developed an automated pipeline for extracting and normalizing ISBNs from the metadata, addressing OCR noise, mixed numeral systems and script complexities that hindered reliable identification. As shown in Figure 1, the pipeline consists of four main components: extraction, normalization, fallback matching, and catalog querying. Given the inconsistencies in the data, it was designed to address the following challenges:

- ISBNs represented using diverse notations (e.g., in Arabic *r-d-m-k*, (رقم الدولي من الكتب), "ISBN", "I.S.B.N")
- The presence of Arabic Indic and Persian numerals in place of Western numerals
- OCR errors, formatting inconsistencies, font variations and missing ISBNs
- Variability in metadata ordering and content across entries.

3.1. Dataset

The dataset for this case study was extracted from the La Pira Library digital collection, a large digital archive of Islamic studies containing diverse documents from multiple institutions [15]. This heterogeneous archive includes approximately 200,000 files totaling nearly 1.2 TB of data. From this

commercial and cataloging purposes.

collection, around 140,000 PDF documents were curated for their cultural, linguistic, and structural variety, ensuring coverage across different subjects (theology, jurisprudence, history, literature), time periods, languages, and visual characteristics (fonts, layouts, page structures). This diversity ensures our evaluation reflects realistic challenges in cultural heritage collections.
To focus on metadata rich content, we built upon a recent study that developed the DigitalMaktaba-LaPira-v1 (DM-LP-v1) dataset, a large scale dataset advancing cataloging and OCR research for multilingual Arabic-script digital libraries [16]. In DM-LP-v1, the first and last ten pages of each document were extracted and processed, as these sections typically contain title pages, colophons, front matter, and bibliographic details such as ISBNs. They were labeled with a 'head'suffix (e.g., `document_name_head.pdf`). The Qwen2-VL-72B Vision Language Model [17] was then applied to identify the most content-rich pages before OCR processing, reducing processing time and storage requirements. This targeted preparation ensured that OCR focused on the most relevant sections and streamlined the workflow.
For this case study[2], we selected one batch yielding 2,470 PDF books. This sample represents a natural cross-section of the collection, including books with varying metadata quality, ISBN formats (Arabic-Indic, Persian, Western numerals), and OCR conditions (from pristine scans to degraded documents). The test set was not pre filtered based on ISBN presence, ensuring realistic evaluation conditions reflecting the challenges of operational deployment in large scale cultural heritage digitization.

3.2. Metadata Extraction from Raw Text

The input files were organized in folder structures containing book-level metadata. Using regular expressions and positional rules (e.g., ISBNs after "ISBN" or ردمك), we extracted: (1) book title, (2) author name, (3) candidate ISBN string, and (4) file location. Extracted values were filtered for numeric patterns (10 or 13 digits) and stored in a structured DataFrame.

3.3. ISBN Normalization and Cleaning

To handle noisy multilingual ISBN representations in Arabic script documents, we developed a normalization component addressing formatting inconsistencies and language specific OCR artifacts. ISBNs extracted from Arabic documents often suffer from directional errors, script mixing, or digit/letter misrecognition.

3.3.1. Transformation and Validation

The first step involves standard cleanup rules applied to all ISBN candidates, regardless of language:

- Prefix Removal: Keywords such as "ISBN," "I.S.B.N," "ر.د.م.ك," and their variations (both Arabic and Latin) are removed from candidate strings. which will create further challenges of digit reversal .
- Character Filtering: Hyphens, colons, punctuation and any non digit characters are removed using regular expressions.
- Numeral Conversion: All Arabic-Indic numerals (U+0660–0669) and Persian (U+06F0–U+06F9) are assigned to ASCII digits (0-9), ensuring uniformity in ISBN validation. This is essential since some texts mix the Western and Arabic numeral systems.
- Length Check: The resulting string must be either 10 or 13 digits long. Shorter candidates are flagged as incomplete and sent to fallback matching (3.4).
- Checksum Validation: Candidates are tested using the official ISBN MOD11 [3] (for 10-digit) and MOD10 [4] (for 13-digit) checksum algorithms. If the candidate fails checksum validation, we attempt character level correction.

[2] https://github.com/saniaaftar/isbn-normalization-pipeline
[3] MOD11 validates ISBN-10 by computing a weighted sum of the first nine digits (weights 10 to 2); if the result modulo 11 is 10, the check digit is 'X'.
[4] MOD10 validates ISBN-13 under the EAN-13 standard by alternately weighting digits by 1 and 3; the check digit ensures the total sum is divisible by 10. See ISBN.org and GS1.org for details.

3.3.2. Error Correction via Character Level

A significant number of ISBN codes extracted using the Qwen-VL Vision Language Model were nearly correct but, in cases where an extracted ISBN candidate fails structural validation (i.e., it contains 10 or 13 digits but fails checksum verification), we apply a targeted error correction mechanism based on character level substitution rules. The confusion dictionary contains common OCR misreadings observed in Arabic script documents. Some examples are listed below (complete dictionary available in our GitHub repository):

- Arabic-Indic "٠" (0) → "٥" or Latin "o"
- Arabic-Indic "٥" (5) → "س" , "٨" (8) → "٥"
- Latin "l" (letter) → "1" (digit)
- Latin "B" → "8"

When an ISBN candidate fails validation but contains one or more characters present in this dictionary, we generate alternate versions by substituting each ambiguous character with plausible digits. Each version is then re validated using the appropriate ISBN checksum. For example, the string 978600100254٥ extracted via Qwen-VL and containing a suspicious final character fails initial validation. Replacing ٥ with 0 produces 9786001002540, which successfully passes MOD10 checksum validation, confirming its validity. This process enables the recovery of ISBNs that would otherwise be discarded due to minor OCR noise. "While prior OCR correction work has focused on Arabic character confusion recovery [**?**], our method extends this by applying checksum based structural validation to recover valid identifiers from partially corrupted strings." This correction mechanism recovered 103 ISBNs (3.4% of successfully extracted ISBNs that initially failed validation).

3.3.3. Handling Bidirectionality and Script Conflicts

Arabic OCR engines frequently misinterpret the directionality of numerals following Arabic text labels (e.g., "ر.د.م.٠ك" ro "رقم دولي"), leading to digit reversal or mixing, since Arabic Indic and Persian digits follow left-to-right reading order, unlike the usual right-to-left. To address this, we analyzed the Unicode bidirectional property of tokens. When an ISBN candidate adjacent to Arabic labels fails checksum validation, we test both the original and reversed digit sequences. If reversal produces a valid checksum, we accept the corrected version. This bidirectionality correction is particularly important when ISBNs appear at page margins or in complex layouts where OCR spatial analysis may fail.

3.3.4. Soft Validation

If an ISBN like string is nearly correct in length (e.g., 11–12 digits) but fails structural validation, it is not discarded. Instead, it is flagged as a soft candidate and passed to the fallback matching process. This maximizes metadata recovery even in degraded conditions.

3.4. Fallback Matching for Incomplete Records

A significant portion of the dataset (over 500 entries) contained missing or malformed ISBNs due to OCR failure, incomplete scans, or manual entry errors. For these records, a fallback matching mechanism was employed using combinations of the book title and author name as secondary keys. This step attempted to identify catalog matches through textual similarity and cross field querying. The fallback mechanism allowed us to continue searching for catalog records even in the absence of valid ISBNs, increasing recall and reducing the likelihood of discarding valuable titles.

3.5. Automated Catalog Querying and Matching

To validate which books were already listed in the "La Pira" physical Library Catalog, we developed a web scraping module using cloudscraper to bypass anti-bot protection and BeautifulSoup [5] to parse

[5] https://pypi.org/project/beautifulsoup4/

HTML content. The matching workflow operated as follows:

1. **Query construction:** For each book record, the system first attempted matching using the `clean_isbn` field. If the ISBN was missing or invalid, it fell back to querying with `book_title` and *author_name*.
2. **Catalog search:** The query was submitted to the La Pira catalog search interface via HTTP request.
3. **Response parsing:** The HTML response was parsed to detect successful matches, identified by the presence of an `<a class="title">` element in the search results.
4. **Result recording:** When a match was found, the catalog URL was stored in `Lapira_url` and a Boolean flag (`Lapira_found = True`) was set. Unmatched records were flagged as `Lapira_found = False`.

To prevent server overload and avoid triggering rate-limiting mechanisms, a 1.5 second delay was introduced between consecutive queries, resulting in a total processing time of approximately 62 minutes for the 2,470 book dataset.

3.6. Baseline Method

As no prior work has reported quantitative ISBN extraction performance on noisy Arabic script OCR data, we developed a regex only baseline for comparison, evaluated on the same 2,470 book records to ensure fair assessment. The baseline approach consists of: (1) pattern matching to extract sequences of exactly 10 or 13 consecutive digits, (2) character filtering to remove non-digits while preserving ”X” as a valid ISBN-10 check digit and (3) checksum validation using MOD11 (ISBN-10) and MOD10 (ISBN-13) algorithms. This baseline deliberately omits numeral system conversion, character level error correction, bidirectionality handling, and fallback matching, thereby isolating the performance gains attributable to our normalization components. Moreover, We deliberately focus on rule based evaluation rather than LLM based comparison because our contribution addresses script specific challenges Arabic Indic numeral conversion, bidirectional text resolution, and OCR confusion pattern correction through specialized engineering. Our target requires zero cost, offline capable solutions, operational constraints incompatible with commercial API dependencies ($ price/token) and internet connectivity requirements. Our baseline thus reflects realistic deployment scenarios for resource constrained cultural heritage digitization rather than comparison with fundamentally different architectural paradigms requiring distinct evaluation frameworks.

Table 1
Performance Metrics for ISBN Extraction

Metric	Baseline Regex based (%)	Pipeline (%)
Precision	61.0	90.4
Recall	58.0	92.5
F1-Score	59.3	87.9
Accuracy	65.0	84.9

4. Discussion

4.1. ISBN Extraction Performance

Our pipeline substantially outperformed the regex only baseline across all metrics (Table 1). Precision improved from 61.0% to 90.4% (+29.4 points), recall from 58.0% to 92.5% (+34.5 points) and F1-score from 59.3% to 87.9% (+28.6 points). These gains validate our hypothesis that Arabic-script ISBN extraction requires specialized normalization beyond simple pattern matching (See Table 1). The pipeline processed

all 2,470 books without prior knowledge of ISBN presence. The extraction process identified 2,000 books containing ISBN candidates, from which 1,850 valid ISBNs were successfully extracted and verified through checksum validation. To validate extraction correctness beyond structural validity, we manually verified a stratified random sample of 200 extracted ISBNs by cross referencing against original PDF scans. This validation confirmed 98% accuracy (196/200 correct extractions), with four errors resulting from page layout ambiguities where multiple ISBNs appeared in different document sections. We additionally inspected all 470 books yielding no ISBN candidates through manual examination of complete document scans, confirming these were either pre-1970 publications (predating ISBN standardization and spreading) or informal publications lacking formal identifiers. Based on this validation, we confirmed 92.5% recall (1,850 correct extractions from 2,000 ISBN bearing books) and 90.4% precision.

To characterize the remaining 150 failed extractions, we conducted error analysis on a stratified random sample of 60 cases. Manual inspection of the original PDF scans and Qwen-identified pages revealed multiple categories such as incomplete or truncated ISBNs (approximately 35% of sampled failures) where only 8-9 digits were extracted due to OCR boundary detection errors, page margin clipping, or partial obstruction; line fragmentation (27%) where ISBN digits were split across multiple text lines during OCR processing; severe OCR degradation (18%) involving extensive digit misrecognition beyond confusion dictionary correction capacity; layout ambiguity (13%) from multiple candidate sequences (ISBN-10/13 variants, barcodes, catalog numbers) that position based heuristics could not disambiguate; uncorrectable bidirectionality errors (5%); and false positive candidates (2%) where non-ISBN numeric sequences matched extraction patterns. Notably, the predominance of truncation and fragmentation issues (62%) suggests these are addressable through enhanced OCR preprocessing and multi line reconstruction in future work. The pipeline generated zero false extractions for the 470 cases without ISBN, demonstrating appropriate precision in negative instances. The performance improvement is attributable to four key components: (1) numeral system conversion, which handles mixed Arabic-Indic, Persian and Western digits; (2) character level error correction using a confusion dictionary derived from common OCR misrecognitions; (3) bidirectionality handling for mixed script sequences; and (4) fallback matching on title and author fields when ISBNs are missing or malformed. The baseline's low performance (61.0% precision) demonstrates that naive regex approaches fail to account for script specific challenges in Arabic OCR data.

4.2. Catalogue Matching and Collection Analysis

Following successful ISBN extraction, we conducted catalog reconciliation against the La Pira Library holdings. We first identified and excluded 726 exact duplicates (29.4%), then employed a two stage matching strategy on the remaining 1,744 unique books. Direct ISBN queries revealed only 7 catalog matches (0.4%), while fallback matching on title and author fields recovered 104 additional entries from the 620 no ISBN cases (16.8% recovery rate), yielding 111 total matches (4.5% overlap).

The analysis identified 1,017 books (61.8% of unique titles) with valid ISBNs absent from the catalog, representing high value acquisition opportunities that underscore the systematic under representation of Arabic script materials in Western library collections. Additionally, 51 books (2.1%) exhibited ISBN conflicts indicating different editions or regional variants, while approximately 400 books (16.2%) lacked sufficient metadata for automated matching and require manual cataloging intervention. These findings reflect broader patterns of underrepresentation for non Western materials in international library systems, suggesting systemic cataloging imbalance beyond technical metadata challenges.

These initial findings demonstrate the pipeline's utility for large scale catalog reconciliation while providing preliminary insights into collection composition. Full scale deployment across the complete archive will yield more robust statistics and enable comprehensive collection development assessment.

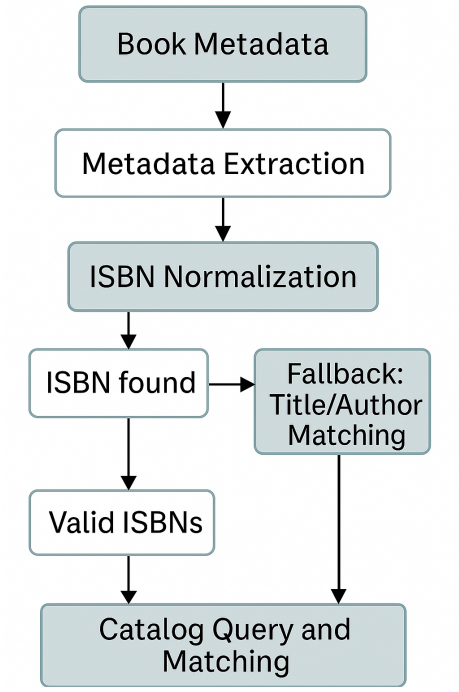

Figure 1: ISBN extraction pipeline

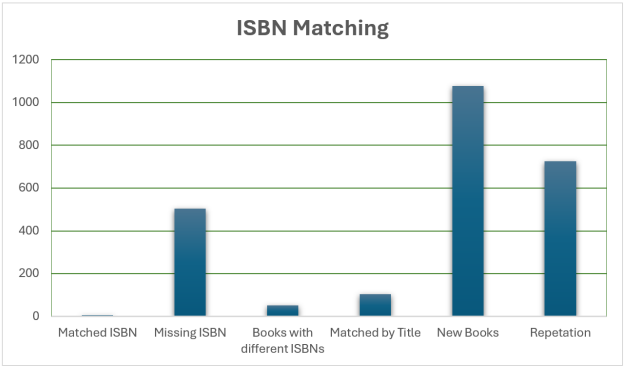

Figure 2: Distribution of ISBN Matching Against La Pira Catalog

4.2.1. Generalizability and Limitations

While our pipeline was developed and evaluated on Arabic script materials, its modular architecture supports adaptation to other non-Latin script contexts. The core normalization components numeral system conversion, character level error correction and bidirectionality handling represent generalizable strategies applicable to Persian (which shares some Arabic Indic numerals but uses additional characters), Urdu and other right-to-left scripts facing similar OCR challenges. For scripts with different numeral systems (e.g., Hindi,Chinese or Hebrew), the confusion dictionary and conversion mappings would require script specific adaptation, but the overall pipeline structure remains applicable.

5. Conclusion

This study introduced an automated pipeline for extracting and normalizing ISBN metadata in Arabic script collections, achieving 90.4% precision and 92.5% recall a 29.4 percentage point improvement over regex only baselines. The pipeline addresses script variability, OCR errors, and formatting inconsistencies through specialized components: numeral system conversion, character level error correction, bidirectionality resolution, and fallback matching. Catalog reconciliation revealed minimal existing overlap (111 books, 4.5%) while identifying 1,078 new titles (43.6%), demonstrating substantial collection development opportunities and underscoring the systematic underrepresentation of Arabic script materials in Western libraries. Manual verification confirmed 98% extraction accuracy with zero false positives among 470 pre ISBN publications. This lightweight, cost free approach proves valuable for resource constrained, script diverse contexts. These findings derive from 2,470 books (1.8% of the collection); statistics may evolve at full scale.

Future work will explore enhanced fuzzy matching, curator interfaces, and scalable deployment strategies to advance cultural heritage digitization while maintaining operational sustainability.

Acknowledgment

This work was conducted within the PNRR project ITSERR - Italian Strengthening of the ESFRI RI RESILIENCE" (Avviso MUR 3264/2022) funded by EU – NextGenerationEU - Grant No IR0000014.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] N. Tscheke, S. Dawson, Book metadata for an open access world: Oa-meta: Open metadata-management for scientific book publications, ScienceOpen Posters (2021).

[2] K. Lybarger, Keeping up with ebooks: Automated normalization and access checking with normac, code4lib Journal (2013).

[3] F. Morillo, I. Santabárbara, J. Aparicio, The automatic normalisation challenge: Detailed addresses identification, Scientometrics 95 (2013) 953–966.

[4] T. J. Skluzacek, K. Chard, I. Foster, Can automated metadata extraction make scientific data more navigable?, in: 2023 IEEE 19th International Conference on e-Science (e-Science), IEEE, 2023, pp. 1–10.

[5] E. Escolano Rodríguez, A. Caro Martín, J. Fejes, I. García-Monge Carretero, M. Gentili-Tedeschi, D. McGarry, R. Ouf, R. Santos Muñoz, H. C. White, I. R. Group, et al., Isbd international standard bibliographic description: 2021 update to the 2011 consolidated edition (2022).

[6] R. N. Westbrook, D. Johnson, K. Carter, A. Lockwood, Metadata clean sweep: a digital library audit project, D-Lib Magazine 18 (2012) 2.

[7] S. Thompson, X. Liu, A. Duran, A. Washington, A case study of etd metadata remediation at the university of houston libraries (2019).

[8] A. Stein, K. J. Applegate, S. R. and, Achieving and maintaining metadata quality: Toward a sustainable workflow for the ideals institutional repository, Cataloging & Classification Quarterly 55 (2017) 644–666. doi:10.1080/01639374.2017.1358786.

[9] P. Király, J. Stiller, V. Charles, W. Bailer, N. Freire, Evaluating data quality in europeana: Metrics for multilinguality, in: Research Conference on Metadata and Semantics Research, Springer, 2018, pp. 199–211.

[10] M. H. Choudhury, L. Salsabil, H. R. Jayanetti, J. Wu, W. A. Ingram, E. A. Fox, Metaenhance: metadata quality improvement for electronic theses and dissertations of university libraries, in: 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2023, pp. 61–65.

[11] V. Christophides, V. Efthymiou, K. Stefanidis, Entity resolution in the web of data, volume 5, Springer, 2015.

[12] G. Chiron, A. Doucet, M. Coustaty, M. Visani, J.-P. Moreux, Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information, in: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017, pp. 1–4. doi:10.1109/JCDL.2017.7991582.

[13] S. Faizullah, M. S. Ayub, S. Hussain, M. A. Khan, A survey of ocr in arabic language: applications, techniques, and challenges, Applied Sciences 13 (2023) 4584.

[14] B. Kiessling, G. Kurin, M. T. Miller, K. Smail, Advances and Limitations in Open Source Arabic-Script OCR: A Case Study, Digital Studies / Le champ numérique 11 (2021). URL: https://www.digitalstudies.org/article/id/8094/. doi:10.16995/dscn.8094, publisher: Open Library of Humanities.

[15] S. Bergamaschi, S. De Nardis, R. Martoglia, F. Ruozzi, L. Sala, M. Vanzini, R. A. Vigliermo, Novel Perspectives for the Management of Multilingual and Multialphabetic Heritages through Automatic Knowledge Extraction: The DigitalMaktaba Approach, Sensors 22 (2022) 3995. URL: https://www.mdpi.com/1424-8220/22/11/3995. doi:10.3390/s22113995, number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[16] L. Sala, R. A. Vigliermo, G. Sullutrone, S. Bergamaschi, DM-LP: A large arabic script languages dataset for OCR and cataloguing research, in: W.-T. Balke, K. Golub, Y. Manolopoulos, K. Stefanidis,

Z. Zhang, T. Aalberg, P. Manghi (Eds.), New Trends in Theory and Practice of Digital Libraries, Springer Nature Switzerland, 2025, pp. 124–134. doi:10.1007/978-3-032-06136-2_12.

[17] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL: https://api.semanticscholar.org/CorpusID:261101015.