Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/150607

A Statistical Characterization of Arabic Historical Manuscript Datasets: Resources, Annotations, Scripts, and Accessibility

Amani Jamal[1,2,*]

[1]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Center for Research Excellence in Artificial Intelligence and Data Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Received: 25 Sep. 2026, Revised: 12 Oct. 2026, Accepted: 20 Oct. 2026
Published online: 1 Nov. 2026

Abstract: Benchmark datasets are the ground on which computational work on Arabic historical manuscripts stands, yet that ground has never been measured. Following PRISMA 2020, we searched Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library (471 records; 27 studies retained), added 8 studies through citation searching, and coded 35 dataset records (2010–2026) twice against their primary papers. Scale is heavily right-skewed (authentic subset: median 300 pages, mean 3,516) and has not grown over time (Spearman $\rho = 0.195$, $p = 0.268$). Fifteen records (42.9%) report a public download, but only eight endpoints delivered open files in October 2026; five (14.3%) name a dataset license and four (11.4%) have a DOI. On a 13-item reporting score, the median record documents 5 items, with no improvement since 2010 ($\rho = 0.13$, $p = 0.44$). Maghrebi hands are the most frequently named script (7 records) yet supply about 1% of authentic pages. Only two records test exclusively on manuscripts unseen during training; where one study reports both protocols, writer-identification accuracy falls from 98.1% to 60.1% (WrIHAM) and from 94.95% to 81% (WAHD) once test manuscripts are kept out of training. We propose a minimum reporting checklist and recommend manuscript-disjoint splits, DOI-backed versioned releases with explicit licenses, and larger Maghrebi collections.

Keywords: applied statistics, PRISMA 2020, systematic review, Arabic historical manuscripts, handwritten text recognition, document image analysis, FAIR data, reporting completeness, document leakage

1 Introduction

Historical Arabic manuscripts hold religious, scientific, legal, literary, and administrative knowledge produced over more than a millennium. Digitization has preserved millions of page images, but images are not yet data. Turning them into searchable evidence requires Document Image Analysis and Recognition (DIAR) systems that find text lines, read cursive handwriting, spot words, identify writers, and assess degradation – all on a script whose letter shapes change with position, whose diacritics fade or drift, and whose pages carry marginalia, bleed-through, and stains [1,2,3].

Such systems are trained and compared on benchmark datasets, and a benchmark is only as informative as its documentation. For historical material one detail matters more than most: pages bound in the same volume share a hand, an ink, a layout, and a vocabulary. If those pages are split at random between training and test, a benchmark measures how well a model reads more of a book it already knows, not how well it reads a new one.

The resources themselves grew task by task rather than from a common plan: word spotting and subword recognition [4,5,6], writer identification [7,8,9,10,11], line and layout analysis [12,13,14,15], Maghrebi-script recognition [16,17], and degradation and quality assessment [18,19,20]. KHATT remains the reference benchmark for modern Arabic handwriting against which historical resources are often contrasted [21]. Existing surveys describe methods and applications [1,3] or compare datasets across many writing systems [2], but none offers a record-level statistical synthesis of the Arabic historical datasets themselves, and none covers the resources released since 2024. Publications and datasets also do not map one to one: a dataset may appear in several versions, and a derived collection may reuse

* Corresponding author e-mail: atjamal@kau.edu.sa

© 2026 NSP
Natural Sciences Publishing Cor.

pages from an earlier one. Without consolidation at the dataset level, the literature overstates how much independent material exists.

This study makes five contributions:

1. A PRISMA 2020 review consolidating 471 database records and 8 citation-identified studies into 35 dataset records (2010–2026), each verified twice against its primary paper, with the page of every extracted fact recorded.
2. A distribution-aware description of scale, coverage, scripts, annotation, and tasks, showing a strongly right-skewed field (authentic median 300 pages, mean 3,516) that has grown in number of datasets but not in their size.
3. A FAIR audit combining what papers report with what their links deliver: 15 records report a public download, but only eight endpoints provided open files in October 2026.
4. A 13-item reporting-completeness score and a matching minimum reporting checklist: the median record reports 5 of 13 essential items, with no improvement over time.
5. An evaluation audit showing that only two records test exclusively on manuscripts unseen during training, and that, where one study reports both protocols, writer-identification accuracy drops by 14 to 38 points when test manuscripts are kept out of training.

Section 2 describes the methods, Section 3 the dataset-level results, Section 4 the evaluation audit, Section 5 the implications and limitations, and Section 6 concludes.

2 Methods

2.1 Review Design and Research Questions

We followed the PRISMA 2020 statement [22]. The unit of analysis is the *dataset record*, not the publication: each retained study contributes one record, anchored in the paper that introduces the dataset. A later paper is kept as a separate record only when it describes a material extension, a distinct version, or a derived resource with its own purpose, and such records are flagged so that they are never counted as independent collections. Ten datasets carry no name in their source paper; we refer to them by author and year (e.g., "the Awaida (2015) dataset") rather than inventing names that could be mistaken for official ones. The two KALIMA releases are distinguished by year. Seven research questions guided the review:

–RQ1. Which Arabic historical manuscript datasets exist for computational research?
–RQ2. How are they distributed in scale, period, geographic origin, domain, and script?
–RQ3. Which annotation types and levels do they provide?
–RQ4. Which tasks do they support, and how do annotation levels map onto tasks?
–RQ5. How accessible, licensed, identifiable, and completely documented are they?
–RQ6. How are benchmarks partitioned, and how many protocols are exposed to manuscript-level leakage?
–RQ7. What gaps remain?

2.2 Search and Selection

We searched Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library (searches completed in September 2026) in title, abstract, and keyword fields with the expression *(Arabic OR "Arabic script") AND (manuscript OR manuscripts OR "historical manuscript" OR "historical manuscripts" OR "historical document" OR "historical documents" OR "handwritten manuscript" OR "handwritten manuscripts") AND (OCR OR HTR OR "handwritten text recognition" OR "optical character recognition" OR "document analysis" OR "layout analysis" OR segmentation OR transcription OR "word spotting" OR "writer identification" OR "script identification")*. The words *dataset*, *database*, or *corpus* were deliberately not required, because many datasets are described inside method papers. Following the PRISMA "other methods" pathway, the reference lists and citing literature of the included studies were then screened.

A study was included when it involved historical documents with a substantial Arabic-script component, introduced or described an identifiable dataset of original or derived historical images, supported at least one computational task (recognition, segmentation, layout analysis, word spotting, writer or script identification, classification, dating, binarization, or degradation analysis), and could be verified from a primary source. We excluded modern-only or printed-only material, humanities studies without a computational dataset, secondary uses of an existing dataset, Arabic-script languages without an eligible Arabic component, and unverifiable resources. Multilingual collections were retained for their Arabic subset.

© 2026 NSP
Natural Sciences Publishing Cor.

The database searches returned 471 records (221 Scopus, 130 Web of Science, 69 IEEE Xplore, 51 ACM). After removing 159 duplicates, 312 records were screened and 283 excluded; of 29 reports sought, one could not be retrieved and one, a proposal without a completed dataset, was excluded at full text, leaving 27 studies. Citation searching added 8 studies, all retained, for a total of 35 records (Fig. 1). Twenty-one records were admitted without restriction ("authentic"), three for an Arabic subset of a multilingual resource, and eleven under an explicit restriction: synthetic or derived data, version overlap, mixed temporal or access scope, an image-quality resource without text, incomplete size, or unverified provenance. Selection, extraction, and coding were performed by a single reviewer, and the protocol was not registered in advance (Section 5.5). A completed PRISMA checklist is supplied as supplementary material.

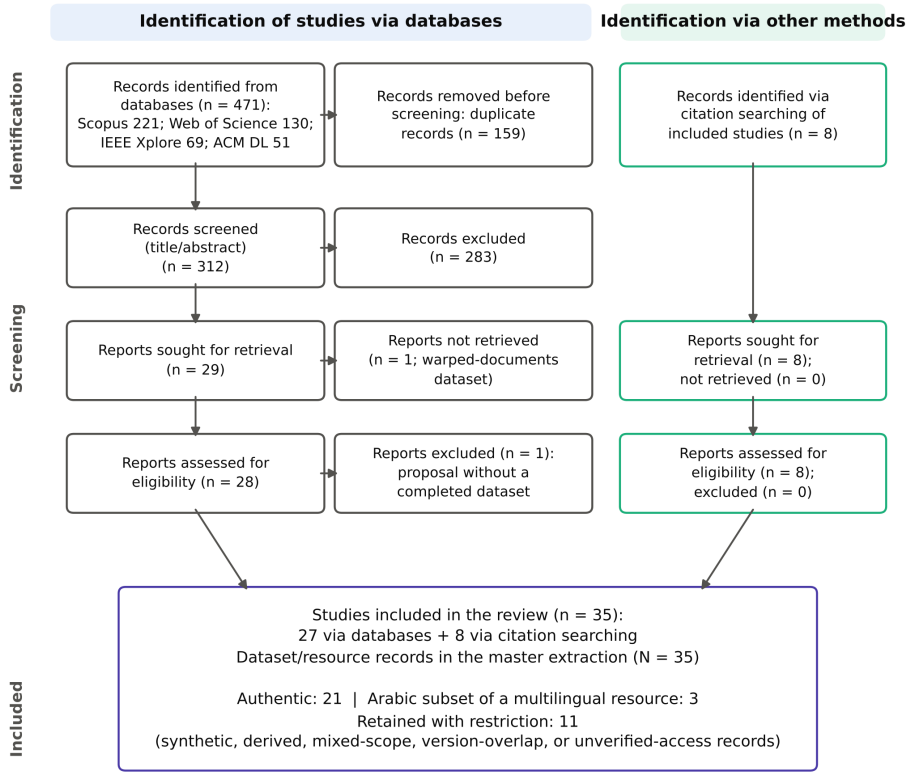


Fig. 1: PRISMA 2020 flow diagram: 471 database records yield 27 studies; citation searching adds 8; the 35 records split into authentic ($n = 21$), Arabic-subset ($n = 3$), and restricted ($n = 11$) strata.

2.3 Extraction, Coding, and Verification

For each record we extracted identification, provenance, period, origin, domain, script, scale (kept in its reported unit), image properties, annotation content and procedure, supported tasks, access, license, persistent identifier, and evaluation protocol. Free-text fields were converted into categorical codes with written rules (annotation level, task, access class, license, DOI, partition, manuscript-disjointness, named styles, image metadata). Two rules deserve mention. A calligraphic style is counted only when the primary paper names it as content of the dataset, never when it is inferred from provenance or mentioned in historical background. And an evaluation is coded *fully manuscript-disjoint* when every reported test manuscript is absent from training, *partially* so when this holds for some experiments or some test material only.

All 35 records were then re-read against their primary PDFs in October 2026. This second pass recorded the page on which each fact appears, added standardized fields, revised several codes (mostly on evaluation protocols and named

© 2026 NSP
Natural Sciences Publishing Cor.

scripts), and logged every change. Internal contradictions in source papers – for instance 95 versus 135 manuscripts in KERTAS, or 98.1% versus 99.1% accuracy in the Balamand paper – are flagged rather than resolved. Dataset endpoints were retrieved on 9 October 2026 to compare reported and actual availability.

Reporting-completeness score. To summarize documentation quality in one number, each record was scored on 13 binary items drawn from the coded fields: precise dates, place of production, domain, named styles, color mode, resolution, file format, annotation quality control, numerical partition, manuscript-disjoint testing (full or partial), public endpoint (full or partial), named dataset license, and dataset DOI. The score counts what a paper *reports*, not whether the reported practice is good.

2.4 Statistical Analysis

Proportions over $N = 35$ are given with Wilson 95% confidence intervals. Because scale is strongly right-skewed, we report medians and interquartile ranges (IQR) alongside means, both for all records with a count and for an *authentic page-level subset* ($n = 29$) that excludes synthetic images, derived lines, word or character crops, and one superseded version. Temporal trends were tested with Spearman's ρ (publication year against $\log_{10}$ pages or against the completeness score) and the Mann–Whitney U test (before 2020 vs. 2020 onward); change in public availability was tested with Fisher's exact test. Tests are two-sided at $\alpha = 0.05$. Benchmark scores were never ranked across studies, because tasks, units, and protocols differ; missing information was coded *not reported* and never imputed.

3 Results

3.1 The Retained Records — RQ1

The 35 records (Tables 1–2) are not 35 independent collections. They include two pairs of successive versions (the 2024 and 2025 KALIMA releases, and RASAM/RASAM 2), a synthetic comparator (SDADDS-Guelma), lines derived from five VML-HD pages, a word dataset that is 89.5% synthetic, 5,240 character crops, three multilingual resources retained for their Arabic subset, and the hierarchical Bouchantouf and Lamghari (2026) dataset, which reuses Muharaf and MOJ-DB pages. Setting aside the superseded version, the derived lines, and the synthetic corpus leaves at most 32 distinct authentic collections, and shared sources (e.g., the Qatar University library behind both MHDID and VDIQA) reduce that number further. Scale values keep the units of their source: manuscripts, pages, photographs, page halves, lines, subwords, crops, and synthetic images are never treated as interchangeable.

3.2 Growth and Scale — RQ2

Twenty-nine records (82.9%) appeared from 2017 onward, but growth is bursty: no record dates from 2011, 2013, or 2016, while 2022 and 2024 contributed five each, and half of all records (18 of 35) appeared in 2022–2026 (Fig. 2). Six of the eight citation-identified studies date from 2020 or later, so recent years are also those where keyword search was least complete.

Scale ranges from 10 pages (VML-ALGN) to 9.6×10^6 synthetic images (SDADDS-Guelma) (Fig. 3). Across the 34 records with a count, the median is 367.5 (IQR 4,854) but the mean is 287,872, almost 800 times larger, because one synthetic corpus dominates; removing it alone brings the mean down to 5,686. Within the authentic page-level subset the skew persists (median 300, IQR 2,441, mean 3,516): WAHD (43,976 pages), the Al-homed et al. (2022) dataset ($\approx$ 14,014), Balamand (11,610), the Awaida (2015) dataset (10,000), and HICMA (5,031) together hold 83.0% of the subset's 101,953 pages and images. Size has not grown with time: Spearman $\rho = 0.195$ ($p = 0.268$) for all records, 0.132 ($p = 0.494$) for the authentic subset, and authentic collections published before and after 2020 have similar medians (275 vs. 300; Mann–Whitney $p = 0.95$). The field produces more datasets, not larger ones.

3.3 Coverage, Origin, and Script — RQ2

Reported coverage runs from the first Islamic century (KERTAS, MHDID, VDIQA) to 2018 (Muharaf), but dates are often coarse. Seventeen records (48.6%) give no precise manuscript dates; KALIMA, for example, lists book years from 632 to 1737 without saying which calendar they use or what they date. Thirteen records (37.1%) do not say where their manuscripts were produced, and thirteen do not report a domain or genre. Named holdings range from the Islamic

© 2026 NSP
Natural Sciences Publishing Cor.

Table 1: Retained Arabic historical manuscript dataset and resource records, in chronological order (Part 1 of 2)

Dataset and year	Reported scale	Historical coverage	Principal computational scope
IBN SINA (2010) [4]	1 manuscript; 51 folios; 20,722 components	Late 12th–early 13th-c. tradition; copy date not reported	Word spotting; subword labeling
Bukhari et al. (2012) dataset† [12]	7 books; 38 page images	Exact dates not reported	Layout analysis (main body vs. side notes)
Fecker et al. (2014) dataset [7]	60 manuscripts; 4,595 pages; 29 writers	13th to early 20th century	Writer identification
HADARA80P (2014) [5]	1 book; 80 images; 16,720 words	833 AH / 1430 CE (single book)	Segmentation-free word spotting
Saabni et al. (2014) dataset [13]	215 pages; at least 166 Arabic pages	Not reported (one 19th-c. subset)	Text-line extraction
Awaida (2015) dataset [8]	200 manuscripts/writers; 10,000 page-half images	Not reported	Writer identification
Sulaiman et al. (2017) dataset [18]	31 degraded document images	Islamic and ancient Arabic eras; exact dates not reported	Degradation analysis; binarization; pre-OCR
VML-HD (2017) [6]	5 books; 680 pages; 121,636 subwords	1088–1451 CE	Subword/word recognition; word spotting
WAHD (2017) [9]	353 manuscripts; 43,976 pages; 322 writer classes (23 known)	13th–20th centuries (IHP)	Writer identification
KERTAS (2018) [23]	95–135 books (inconsistent in paper); 2,505 images	Fourteen Islamic centuries	Manuscript dating
MHDID (2018) [19]	130 books; 335 images	1st–14th Islamic centuries	Image-quality assessment; degradation classification
RASM2018 (2018) [14]	100 pages (15 development, 85 evaluation)	8th–19th centuries	Page segmentation; line detection; OCR
VDIQA† (2018) [24]	177 full-page images	1st–14th centuries AH	Image-quality assessment
BADAM (2019) [15]	42 manuscripts; 400 pages; 10,770 baselines	Dates not reported	Baseline detection; layout analysis
Alsaedi et al. (2020) datasets† [25]	5,240 character crops in three sets	7th–8th Hijri centuries	Character recognition
Maghrebian Font Database† (2020) [26]	20 original images (+20 ground truth); 22,000 component images; 5,000 words	Partly 13th century; full range not reported	Binarization; enhancement
RASAM (2021) [16]	3 manuscripts; 300 images; 7,540 lines	1734–1875 CE	HTR; layout and baseline analysis
Al-homed et al. (2022) dataset [27]	108 manuscripts; approx. 14,014 images	Exact dataset dates not reported	Manuscript classification; retrieval; dating fields

†Identified through citation searching (PRISMA "other methods" branch).

Table 2: Retained Arabic historical manuscript dataset and resource records, in chronological order (Part 2 of 2)

Dataset and year	Reported scale	Historical coverage	Principal computational scope
Balamand Arabic Historical Dataset (2022) [10]	567 manuscripts; 11,610 photographs	13th–19th c. (mostly 17th–19th); 158 undated	Writer identification
MOJ-DB (2022) [28]	10 books; 64 pages; 560,000 subword samples	16th–17th centuries	Subword extraction and recognition; word spotting
VML-AHTE† (2022) [29]	30 pages (20 train, 10 test)	Exact dates not reported	Text-line extraction
VML-ALGN (2022) [30]	2 manuscript copies; 10 pages	Copies of a work authored ca. 1127–1128 CE	Manuscript/text-line alignment; subword detection
HICMA (2023) [32]	5,031 line images (3,434 from manuscripts)	11th and 17th centuries; web calligraphy undated	OCR/HTR; calligraphic-style identification
Maouloud et al. (2023) dataset† [31]	3 manuscripts; 30,430 word images (3,196 real, 27,234 synthetic)	Copy dates not reported	Word recognition (HTR)
Historical Arabic Handwritten Text Recognition Dataset (2024) [33]	8 books; 40 pages; 640 lines	Different eras; exact dates not reported	OCR/HTR
KALIMA (2024 release)† (2024) [34]	2 British Library manuscripts; 26 pages; 497 lines	Exact dates not reported	Word detection and recognition
Maouloud and Ba (2024) dataset [36]	50,000 synthetic lines from five VML-HD source pages	Inherited from VML-HD (1088–1451 CE)	Text-line HTR
Muharaf (2024) [35]	1,644 pages; 36,311 lines	1800–2018	HTR; segmentation; binarization; font and writer recognition
RASAM 2† (2024) [17]	15 manuscripts; 250 images; 3,750 transcribed lines	Dates not reported	HTR; layout analysis
KALIMA (2025) [37]	10 books; 116 pages; 2,431 lines	Book years 632–1737 (calendar not stated); two undated	Text-line detection; HTR/OCR
SDADDS-Guelma (2025) [20]	>9.6 million synthetic degraded images	Synthetic; no intrinsic manuscript date	OCR/HTR; segmentation; binarization; restoration
Bouchantouf and Lamghari (2026) dataset [38]	Total size not reported	Not reported; reuses Muharaf (1800–2018)	Line/word/pseudo-word/character segmentation
MIRATH (2026) [39]	3,000 pages	9th–19th centuries	Text-line segmentation; layout analysis
OpenITI MAKHZAN (2026) [40]	208 documents; 1,497 pages; 822 Arabic pages	900–1990 CE	HTR/OCR
WrIHAM (2026) [11]	62 books; 1,144 pages; 18 writers	11th–17th centuries CE	Writer identification/verification

†Identified through citation searching (PRISMA "other methods" branch).

© 2026 NSP
Natural Sciences Publishing Cor.

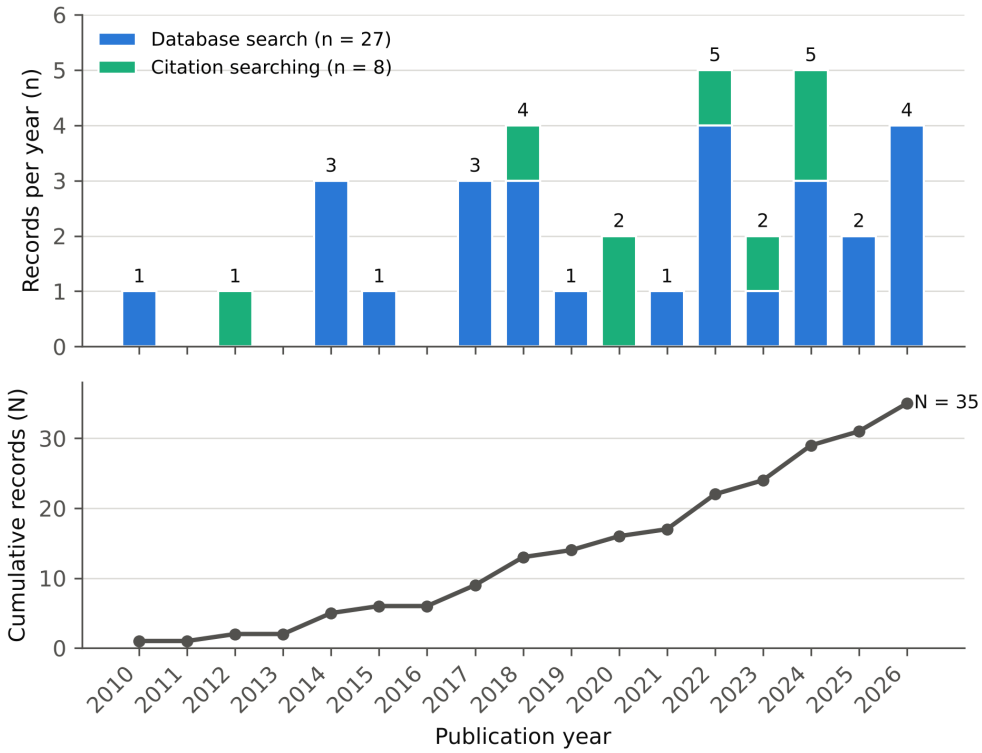


Fig. 2: Records per year by identification route (top) and cumulative count (bottom), 2010–2026.

Heritage Project, the Qatar National and Digital Libraries, and the Prophet's Mosque Library to BULAC, Leipzig, the Chester Beatty Library, the Tunisian National Archives, Endangered Archives Programme collections from West Africa, and Antiochian Orthodox repositories in Lebanon and Syria.

Only ten records (28.6%) name the calligraphic styles of their data. Three plausible codings failed the naming rule on re-reading: KERTAS mentions Kufic, Thuluth, and Naskh only in its historical background; MOJ-DB comes from Fez but never names its script; and the "Maghrebi" of SDADDS-Guelma is a printed font used for synthesis. Among the ten, the Maghrebi family appears in seven (20.0% of all records), ahead of Naskh (5), Thuluth, Kufic, and Farisi/Nastaliq (3 each), with single mentions of Ruq'ah, Diwani, Muhaqqaq, Ta'liq, and Shikaste. Four resources are entirely Maghrebi (RASAM, RASAM 2, the Maghrebian Font Database, and the Al-Soudani word dataset of Maouloud et al. (2023)). By records, Maghrebi script is the best-labelled tradition in the field.

By pages, the picture flips. RASAM (300), RASAM 2 (250), the Maghrebian Font Database (20), and the 16% Maghribi share of MIRATH ($\approx$ 480) add up to about 1,050 pages, or 1.0% of authentic pages. This is a lower bound – OpenITI MAKHZAN and the Bouchantouf and Lamghari (2026) dataset give no Maghribi counts, and MOJ-DB's 64 pages may be Maghribi – but including all of them would lift the share only to 1.1%. Maghrebi collections are not unusually small (median 300 pages vs. 275 for the others); the issue is that none of the five largest collections, holding 83% of authentic pages, contains labelled Maghrebi material. The *Maghrebi Script Gap* is thus a gap of volume, not of presence.

3.4 Image Metadata — RQ2

Eighteen records (51.4%) state a color mode, thirteen (37.1%) a resolution or pixel size, and eleven (31.4%) a file format; twelve (34.3%) report none of the three. The seven records that state a DPI value range from 96 to 500 DPI, and BADAM alone mixes 200, 300, and 500 DPI scans. Such differences are almost never controlled for when results are compared across datasets.

© 2026 NSP
Natural Sciences Publishing Cor.

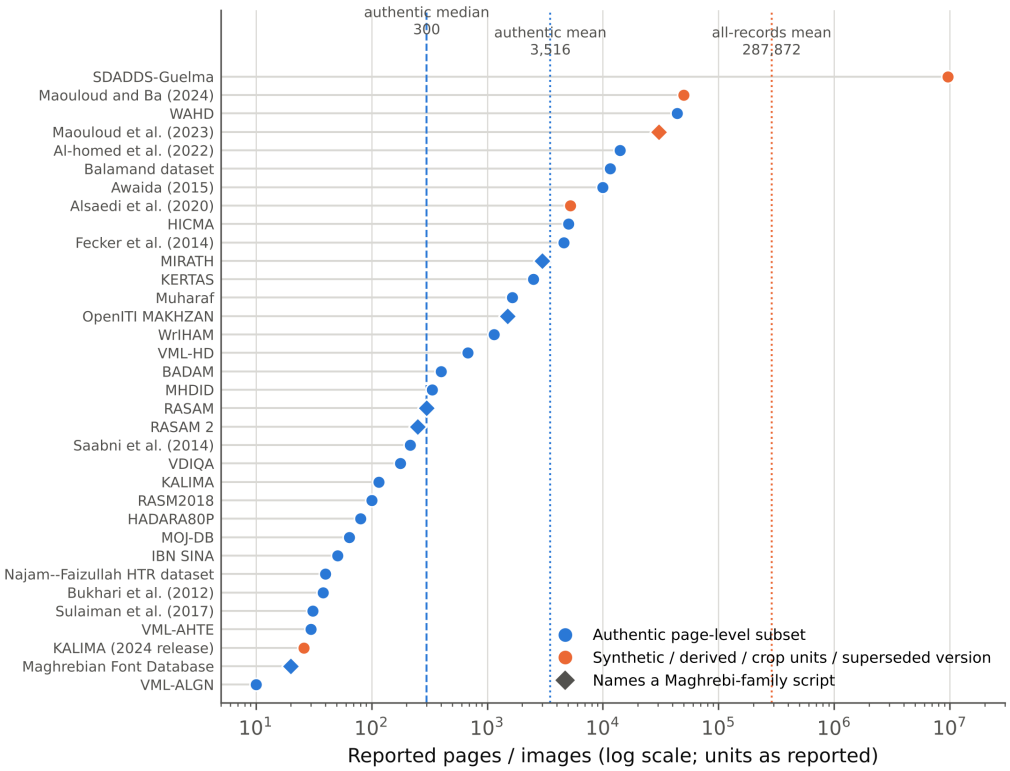


Fig. 3: Reported scale of the 34 records with a page/image count (log scale, units as reported). Blue: authentic page-level subset; orange: synthetic, derived, crop-unit, or superseded records; diamonds: records naming a Maghrebi-family script. Lines mark the authentic median and mean and the all-records mean.

3.5 Annotation and Tasks — RQ3/RQ4

Annotation ranges from page labels to character targets (Table 3). Line or baseline ground truth (16 records) and transcriptions (18) are common; character-level labels are rare (5, 14.3%), and only six records (17.1%) combine three or more spatial levels. Text recognition is the most frequent task (16 records, 45.7%), followed by line or baseline segmentation (10), layout analysis (6), writer identification (5), word spotting (4), restoration or binarization (4), quality assessment (3), script classification (2), date or document classification (2), and cross-copy alignment (1). The contingency matrix (Fig. 4) shows line annotation as the workhorse of both recognition (11) and segmentation (10), writer identification resting entirely on page-level labels, and character-level annotation nearly absent outside recognition. Formats are fragmented: PAGE XML (seven records), ALTO XML, Hadara XML, JSON, CSV, MATLAB structures, Labelme rectangles, and YOLO/COCO/VOC exports all occur.

3.6 Access, Licensing, and Persistence — RQ5

Fifteen records (42.9%; 95% CI 28.0–59.1%) are reported as publicly downloadable, ten (28.6%; 16.3–45.1%) give no verifiable endpoint, six (17.1%) only promise a release, and four are partial, on request, or restricted by the holding library (Fig. 5, Table 4). We classify by what the paper says: HADARA80P, MOJ-DB, and IBN SINA count as public because they print an endpoint, while KALIMA and VML-ALGN count as planned because they phrase release in the future tense. Openness has not improved: 8 of 14 records before 2020 and 7 of 21 since are reported public (Fisher $p = 0.19$).

Reported access is an upper bound. Of the 20 records with a recorded endpoint, only eight delivered open files when checked; three offered data on request, five redirected to a generic page after personal or laboratory sites were retired, and four could not be resolved (Table 5). Only six of the 15 records reported as public could be confirmed. Deposits and papers also disagree: BADAM's Zenodo record lists CC BY-SA 4.0 rather than the 2.0 in the paper, SDADDS-Guelma's deposit declares a license the paper omits, and two DOIs exist that the papers never mention. Only five records (14.3%;

© 2026 NSP
Natural Sciences Publishing Cor.

Table 3: Annotation categories across the 35 retained records (a record may belong to several categories)

Annotation category	*n* (%)	Dataset records
Page- or image-level labels	16 (45.7%)	Fecker et al. (2014) dataset; HADARA80P; Awaida (2015) dataset; Sulaiman et al. (2017) dataset; WAHD; KERTAS; MHDID; VDIQA; BADAM; Al-homed et al. (2022) dataset; Balamand Arabic Historical Dataset; HICMA; Historical Arabic Handwritten Text Recognition Dataset; Muharaf; KALIMA; WrIHAM
Layout or region annotation	10 (28.6%)	Bukhari et al. (2012) dataset; HADARA80P; RASM2018; BADAM; RASAM; Muharaf; RASAM 2; KALIMA; SDADDS-Guelma; MIRATH
Baseline or line annotation	16 (45.7%)	Saabni et al. (2014) dataset; RASM2018; BADAM; RASAM; MOJ-DB; VML-AHTE; Historical Arabic Handwritten Text Recognition Dataset; KALIMA (2024 release); Maouloud and Ba (2024) dataset; Muharaf; RASAM 2; KALIMA; SDADDS-Guelma; Bouchantouf and Lamghari (2026) dataset; MIRATH; OpenITI MAKHZAN
Word, subword, or connected-component annotation	11 (31.4%)	IBN SINA; HADARA80P; VML-HD; Maghrebian Font Database; MOJ-DB; VML-ALGN; Maouloud et al. (2023) dataset; KALIMA (2024 release); KALIMA; SDADDS-Guelma; Bouchantouf and Lamghari (2026) dataset
Character-level labels or targets	5 (14.3%)	IBN SINA; VML-HD; Alsaedi et al. (2020) datasets; KALIMA; Bouchantouf and Lamghari (2026) dataset
Transcription or textual labels	18 (51.4%)	IBN SINA; HADARA80P; VML-HD; RASM2018; Maghrebian Font Database; RASAM; MOJ-DB; VML-ALGN; HICMA; Maouloud et al. (2023) dataset; Historical Arabic Handwritten Text Recognition Dataset; KALIMA (2024 release); Maouloud and Ba (2024) dataset; Muharaf; RASAM 2; KALIMA; SDADDS-Guelma; OpenITI MAKHZAN
Dedicated writer or scribe labels	5 (14.3%)	Fecker et al. (2014) dataset; Awaida (2015) dataset; WAHD; Balamand Arabic Historical Dataset; WrIHAM
Degradation or quality labels	4 (11.4%)	Sulaiman et al. (2017) dataset; MHDID; VDIQA; SDADDS-Guelma
Descriptive manuscript metadata	9 (25.7%)	Fecker et al. (2014) dataset; HADARA80P; WAHD; KERTAS; Al-homed et al. (2022) dataset; Balamand Arabic Historical Dataset; MIRATH; OpenITI MAKHZAN; WrIHAM

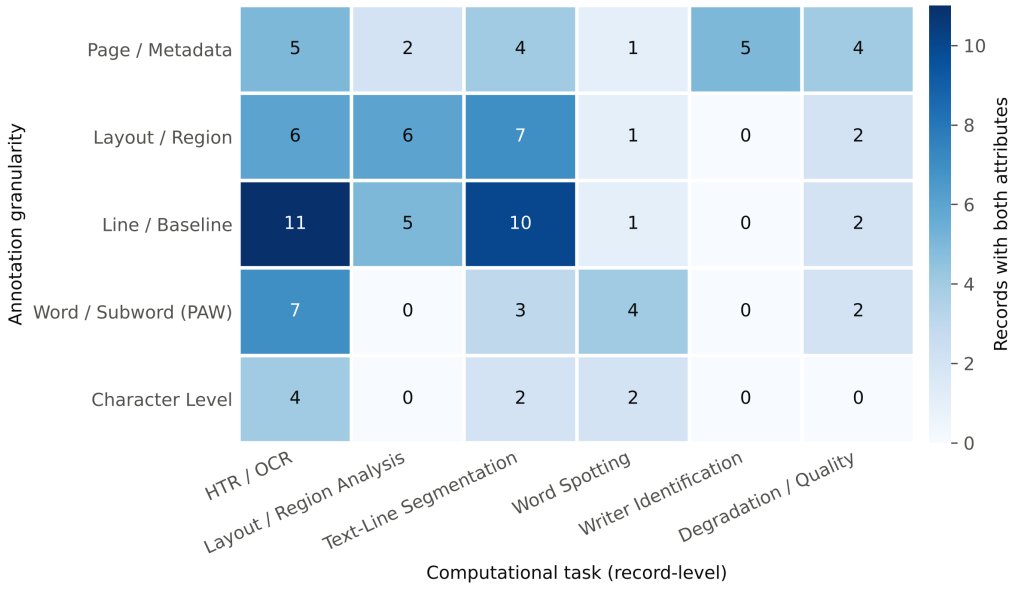


Fig. 4: Annotation granularity against computational task (cells count records with both attributes).

Table 4: Primary access classification of the 35 retained records

Access class	*n* (%)	Dataset records
Reported publicly downloadable	15 (42.9%)	IBN SINA; HADARA80P; Saabni et al. (2014) dataset; VML-HD; WAHD; RASM2018; VDIQA; BADAM; RASAM; MOJ-DB; VML-AHTE; HICMA; Historical Arabic Handwritten Text Recognition Dataset; RASAM 2; OpenITI MAKHZAN
Partially public	2 (5.7%)	Muharaf; SDADDS-Guelma
Available on request	1 (2.9%)	Bouchantouf and Lamghari (2026) dataset
Restricted by source institution	1 (2.9%)	WrIHAM
Planned for release	6 (17.1%)	KERTAS; Balamand Arabic Historical Dataset; VML-ALGN; KALIMA (2024 release); KALIMA; MIRATH
Unverified or unavailable	10 (28.6%)	Bukhari et al. (2012) dataset; Fecker et al. (2014) dataset; Awaida (2015) dataset; Sulaiman et al. (2017) dataset; MHDID; Alsaedi et al. (2020) datasets; Maghrebian Font Database; Al-homed et al. (2022) dataset; Maouloud et al. (2023) dataset; Maouloud and Ba (2024) dataset

6.3–29.4%) name a dataset license, and four (11.4%; 4.5–26.0%) have a dataset DOI. Article licenses were never counted as dataset licenses.

© 2026 NSP
Natural Sciences Publishing Cor.

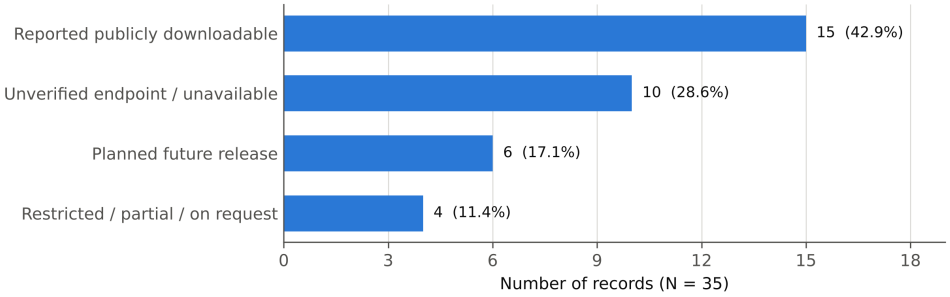


Fig. 5: Access as reported in the primary papers ($N = 35$).

Table 5: Verification of the recorded dataset endpoints (9 October 2026)

Outcome	*n*	**Records**
Open download confirmed	8	SDADDS-Guelma; OpenITI MAKHZAN; BADAM; Muharaf (public part); HICMA; RASAM; RASAM 2; Najam–Faizullah HTR dataset
Online, data on request	3	RASM2018 (images); MOJ-DB; IBN SINA
Redirect to generic page	5	WAHD; VML-HD; Fecker et al. (2014); Saabni et al. (2014); HADARA80P
Not resolvable	4	WrIHAM; KALIMA; VDIQA; VML-AHTE

Note: "Not resolvable" means the retrieval tool could not reach the resource; it does not establish that the endpoint is broken.

3.7 Reporting Completeness — RQ5/RQ7

The 13-item score condenses the gaps above. The median record reports 5 of 13 items (IQR 3.5–8; range 1–11; mean 5.5). Only RASAM (11) and OpenITI MAKHZAN (10) report ten or more, while the Awaida (2015) dataset and the 2024 KALIMA release report one each (Fig. 6). Documentation has not improved over time: the score is unrelated to publication year (Spearman $\rho = 0.13$, $p = 0.44$), and records published before and after 2020 have similar medians (5.5 vs. 5; Mann–Whitney $p = 0.88$). Item by item (Table 6), the best-reported facts are a numerical partition (68.6%), place of production, and domain (62.9% each); the worst are manuscript-disjoint testing (22.9%), a named license (14.3%), and a DOI (11.4%). Annotation quality control shows how far description stops short of measurement: 21 records (60.0%) describe some form of checking, but only one – the Al-homed et al. (2022) dataset, with Cohen's kappa – reports an agreement statistic, and MIRATH, which double-annotated 10% of its pages, does not give the resulting value.

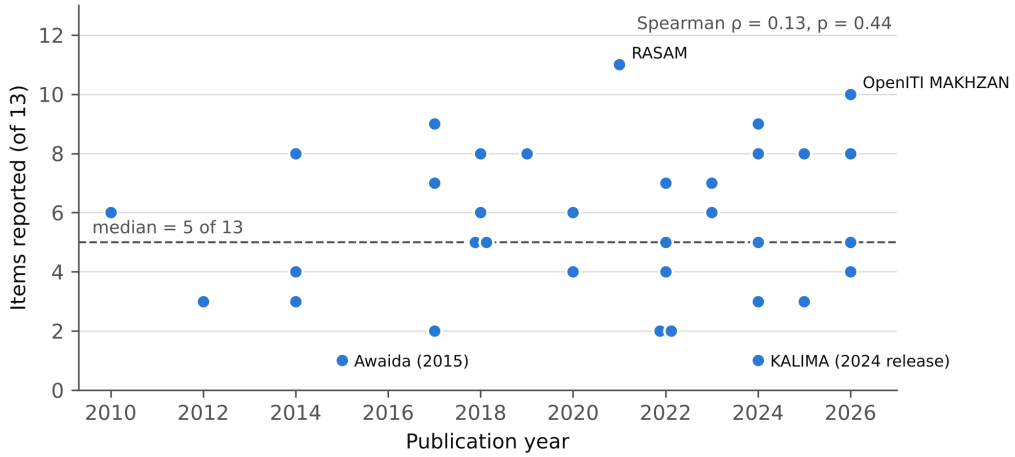


Fig. 6: Reporting-completeness score (items reported out of 13) by publication year. Each dot is one record; the dashed line is the median.

© 2026 NSP
Natural Sciences Publishing Cor.

Table 6: Minimum reporting checklist for Arabic historical manuscript datasets, with the share of the 35 records that currently report each item

Item	What to report	*n* (%)
Provenance		
Dates	Copy date or century per manuscript, calendar stated	18 (51.4)
Origin	Place of production, not only holding library	22 (62.9)
Domain	Subject or genre per manuscript	22 (62.9)
Script	Named styles, from the documents themselves	10 (28.6)
Images		
Color mode	Color, grayscale, or binary, as released	18 (51.4)
Resolution	DPI or pixel dimensions	13 (37.1)
File format	Format of released images	11 (31.4)
Annotation		
Quality control	Who checked, how, and an agreement statistic	21 (60.0)
Evaluation		
Partition	Train/validation/test counts and a split file	24 (68.6)
Disjointness	Whether test manuscripts are unseen in training	8 (22.9)
Release		
Endpoint	A working download location	17 (48.6)
License	A named license for images and annotations	5 (14.3)
Identifier	A dataset DOI with version	4 (11.4)

Note: Counts are items *reported*; "Endpoint" includes partially public records, "Disjointness" includes partial designs, and "Quality control" counts any described check (only one record reports an agreement statistic).

4 Benchmark Design and Document Leakage — RQ6

Twenty-four records (68.6%; 95% CI 52.0–81.4%) report a numerical partition: 2,400/300/300 pages for MIRATH, 320/80 for BADAM, 20/10 for VML-AHTE, 80/10/10 for HICMA, 70/20/10 for MOJ-DB, 15 development plus 85 evaluation pages for RASM2018, and so on. Others use task-specific protocols such as HADARA80P's 25 query keywords or MHDID's subjective ratings.

Reporting a split is not the same as making it independent. For each record we asked one question: does any test manuscript also appear in training? Only two records can answer no for every experiment (5.7%; 95% CI 1.6–18.6%): WrIHAM assigns whole books to training, validation, and test, and Fecker et al. hold out each manuscript of a multi-manuscript writer in turn. Six records (17.1%) do so for part of their evidence – WAHD through an additional leave-one-manuscript-out experiment, Balamand through a test set "a large part" of which is unseen, RASAM through cross-manuscript tests, RASAM 2 and OpenITI MAKHZAN through out-of-domain sets, and Bukhari et al. through test pages "from different books." The remaining 27 (77.1%; 95% CI 61.0–87.9%) use random page-, line-, or sample-level partitions, or none.

This matters because a random split shows a model the hand, ink, ruling, vocabulary, and scanner of each test manuscript during training. Derived and synthetic data make the overlap unavoidable: the VML-HD-derived lines come from five pages, and most samples of the Maouloud et al. (2023) dataset and of MOJ-DB are synthetic variants of a few thousand originals. Two protocols go further: the Maghrebian Font Database reports its F-measure of 98.1% on the very images its training windows were cut from, and HICMA scores its zero-shot baselines on the validation subset. Reported error rates also span two orders of magnitude – from a CER of 1.31% on synthetic VML-HD lines to a WER of 94.51% for zero-shot EasyOCR on HICMA – reflecting data and protocol at least as much as models, which is why we do not rank them.

4.1 Writer Identification: What Happens When the Manuscript Changes

Five writer-identification datasets share a task and a metric (top-1 accuracy), which makes them the clearest place to see what partition design does (Table 7, Fig. 7).

The strongest evidence comes from inside two studies, where the corpus stays fixed and only the split changes. WrIHAM's best model reaches 98.1% when pages of each test book were seen in training and 60.1% on entirely new books: 38.0 points lost to nothing but the position of the train–test boundary. WAHD moves the same way, from 94.95% with a page-level split to 81% with held-out manuscripts (−13.95 points), although its two experiments also use different feature pipelines.

Across studies the picture is less tidy. Fecker et al. test on unseen manuscripts and still report 92.5% (100% at manuscript level), so disjoint testing does not doom a dataset to low scores; their task, with 11 multi-manuscript writers

© 2026 NSP
Natural Sciences Publishing Cor.

Table 7: Writer-identification protocols and reported top-1 accuracy

Dataset (year)	Mss./ writer	Protocol as reported	Seen mss.	Unseen mss.
Awaida (2015)	1.00	8,000/2,000 page halves; one manuscript per writer	93.95%	–
WAHD (2017)	1.10	Page split (GMMS) vs. leave-one-manuscript-out (OBI/SIFT)	94.95%	81%
Fecker et al. (2014)	2.07	Leave-one-manuscript-out	–	92.5%[a]
Balamand (2022)	1.29‡	3,110/8,500 images; test partly from unseen manuscripts	99.1%[b] (mixed)	
WrIHAM (2026)	3.44	Book-shared vs. book-disjoint, same data and models	98.1%	60.1%

Note: Seen mss. = test pages come from manuscripts also used in training; Unseen mss. = every test manuscript is absent from training. [a]Page level; 100% at manuscript level. [b]Table 7 of the source; its abstract reports 98.1%. ‡Uses the 329 manuscripts with an identified copyist (2.21 with all 567).

and a homogeneous scanning pipeline, is simply easier than WrIHAM's 18 writers across 62 books from libraries in eight countries. Partition design, task difficulty, and corpus homogeneity all move the number, so we do not pool accuracies across papers. The within-study contrasts are the evidence, and they agree on direction.

The manuscript-to-writer ratio explains why the cleaner test is rare. Pooled over the five resources it is 1.22 (1,004 manuscripts for 825 writers). With one manuscript per writer, the 93.95% of the Awaida (2015) dataset is necessarily obtained on manuscripts seen in training, and WAHD can hold out manuscripts for only 11 of its 23 known writers. Only WrIHAM, with at least three books per writer, lets every writer be tested on an unseen book.

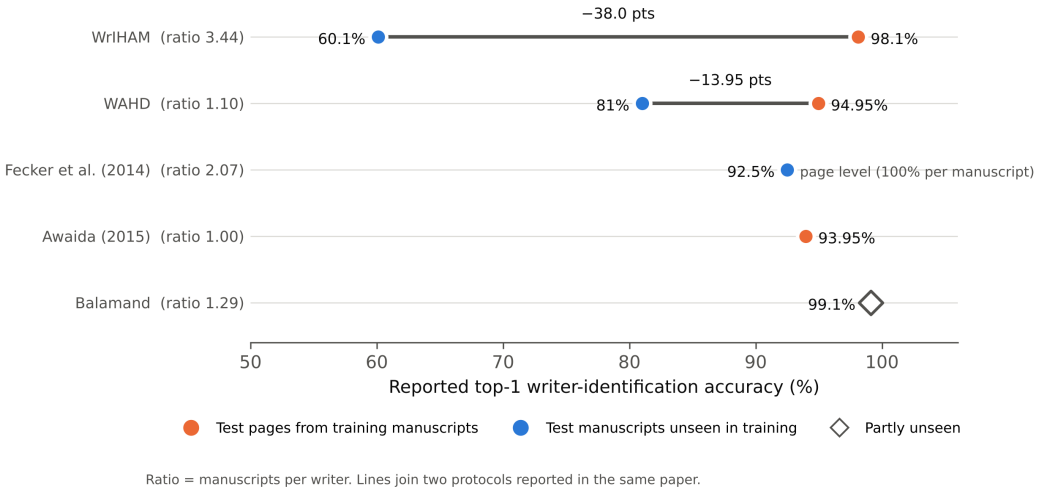


Fig. 7: Top-1 writer-identification accuracy by protocol. Orange: test pages from training manuscripts; blue: unseen test manuscripts; open diamond: partly unseen. Lines join two protocols from the same paper (WrIHAM −38.0, WAHD −13.95 points).

5 Discussion

5.1 What the Evidence Shows

The 35 records cover recognition, segmentation, writer analysis, word spotting, dating, binarization, and quality assessment, yet they do not form a benchmark ecosystem. Three findings stand out. First, the field has grown in the

© 2026 NSP
Natural Sciences Publishing Cor.

number of datasets but not in their size, and a handful of large writer-identification and classification collections dominate the page count. Second, documentation is thin and static: the median record reports five of thirteen essential items, a figure that has not moved since 2010, and openness is claimed more often than delivered. Third, most published scores answer a narrower question than the one usually implied – how well a model reads more pages of a familiar manuscript – and where the wider question has been asked inside one study, the answer is markedly lower. The eight studies that only citation searching could find (22.9% of the final set) add a fourth, quieter finding: much of the field's data lives inside method papers rather than in discoverable dataset publications.

5.2 Representativeness and the Maghrebi Gap

Coverage spans more than a millennium, but nearly half the records cannot place their manuscripts in time and over a third cannot place them in space; the observable distribution reflects which collections were digitized and accessible, not the surviving manuscript heritage. The Maghrebi case shows why records and pages must be counted separately. Maghrebi hands now have dedicated, well-documented resources – RASAM is the most completely reported dataset in the review – but their combined volume is about one percent of authentic pages, which caps the training data available for North African, Andalusian, and West African archives. Diwani, Muhaqqaq, Ruq'ah, and Ta'liq, named in one record each, are in a weaker position still. Script labels must come from the documents themselves: a Fez provenance or a suggestive acronym does not say which hand a page is written in, and MOJ-DB shows how easily such an inference could slip into a statistic.

5.3 From Availability to Reusability

A dataset described as available is not necessarily reusable. Personal and laboratory pages were the least durable channel in our check, license statements differed between papers and deposits, and article licenses were sometimes the only license in sight – although they govern the article, not manuscript images that libraries may still control. Image metadata compounds the problem: when a third of records report neither color mode, resolution, nor format, nobody can tell whether a score reflects the model or an unstated difference in acquisition. The checklist in Table 6 turns these gaps into thirteen concrete items. None requires new data, only that existing knowledge be written down; its weakest items (license, DOI, disjointness) are also the cheapest to fix at release time.

5.4 Evaluation Practice

For writer identification, the within-study contrasts put a price on document leakage of 14 to 38 points. We cannot say how large the effect is for text recognition or layout analysis, because no paper in our set reports both protocols on the same recognition or segmentation data; that experiment is cheap and overdue. The Fecker et al. result is a useful counterweight: high accuracy on unseen manuscripts is possible when the task allows it, so the point is not that disjoint scores must be low but that only disjoint scores answer the generalization question. Benchmarks should therefore report within-manuscript and manuscript-disjoint performance side by side, publish split files grouped by manuscript, report the manuscript-to-writer (or manuscript-to-class) ratio, and keep synthetic and derived samples on one side of the split as their sources. The same applies with greater force to vision-language models [41,42,43]: large pretrained models can memorize images, so a leakage-prone benchmark would reward recall rather than paleographic generalization. Their arrival raises the value of interoperable, multi-level annotation (PAGE-XML or JSON-LD linking page, region, line, word, and character), but it does not remove the need for manuscript-disjoint ground truth.

5.5 Limitations

First, the search covered four bibliographic databases; citation searching added a quarter of the final set, so other resources released only through repositories or inside method papers may have been missed. Second, selection, extraction, and coding were done by one reviewer without a registered protocol, so no inter-rater agreement can be reported; the second verification pass, the page-level evidence, and the published coding rules mitigate but do not replace double coding. Third, inclusion of mixed, derived, versioned, and synthetic resources required judgment, which we labelled and analyzed separately. Fourth, access status is a snapshot of October 2026, and four endpoints could not be resolved. Fifth, several analyses rest on small numbers – five writer-identification datasets, seven Maghrebi-labelled records, two within-study contrasts – so non-significant results are not evidence of absence, and the completeness score weights all thirteen items equally and measures reporting, not quality.

© 2026 NSP
Natural Sciences Publishing Cor.

6 Conclusion

This review characterized 35 Arabic historical manuscript dataset records published between 2010 and 2026, each checked twice against its primary paper. The field is productive but not cumulative. Scale is heavily skewed and has not grown; only 42.9% of records report a public download, eight endpoints actually delivered open files, 14.3% name a dataset license, and 11.4% have a DOI; the median record reports five of thirteen essential items, with no improvement over time. Maghrebi script is the most frequently named tradition yet supplies about 1% of authentic pages. And only two records test exclusively on manuscripts unseen in training, while the two studies that compare protocols lose 14 to 38 points of writer-identification accuracy when they do.

The remedies are modest and mostly a matter of practice: release datasets with a DOI, a version, and a named license for images and annotations; document provenance, script, and acquisition using a checklist such as Table 6; distribute manuscript-grouped split files and report disjoint results alongside within-manuscript ones; keep synthetic data separate from authentic evaluation; and grow Maghrebi and other under-represented collections in pages, not only in number. These steps would let future results be compared, reproduced, and trusted.

Declarations

Funding

This research received no external funding.

Competing interests

The author declares that there are no competing interests.

References

[1] M. Ibn Khedher, H. Jmila and M.A. El-Yacoubi, Pattern Recognition **100**, 107144 (2020).
[2] K. Nikolaidou, M. Seuret, H. Mokayed and M. Liwicki, International Journal on Document Analysis and Recognition **25**, 305-338 (2022).
[3] S. Faizullah, M.S. Ayub, S. Hussain and M.A. Khan, Applied Sciences **13**, 4584 (2023).
[4] R. Farrahi Moghaddam, M. Cheriet, M.M. Adankon, K. Filonenko and R. Wisnovsky, IBN SINA: a database for research on processing and understanding of Arabic manuscripts images, In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 11-18 (2010).
[5] W. Pantke, M. Dennhardt, D. Fecker, V. Märgner and T. Fingscheidt, An historical handwritten Arabic dataset for segmentation-free word spotting – HADARA80P, In: Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), 15-20 (2014).
[6] M. Kassis, A. Abdalhaleem, A. Droby, R. Alaasam and J. El-Sana, VML-HD: the historical Arabic documents dataset for recognition systems, In: Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), 11-14 (2017).
[7] D. Fecker, A. Asi, V. Märgner, J. El-Sana and T. Fingscheidt, Writer identification for historical Arabic documents, In: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR), 3050-3055 (2014).
[8] S.M. Awaida, Text independent writer identification of Arabic manuscripts and the effects of writers increase, In: Proceedings of the International Conference on Computer Vision and Image Analysis Applications, 1-4 (2015).
[9] A. Abdelhaleem, A. Droby, A. Asi, M. Kassis, R. Al Asam and J. El-Sana, WAHD: a database for writer identification of Arabic historical documents, In: Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), 64-68 (2017).
[10] M. Chammas, A. Makhoul, J. Demerjian and E. Dannaoui, Multimedia Tools and Applications **81**, 30769-30784 (2022).
[11] A. Saleh, W.G. Al-Khatib, M. Amro and A. Almugren, IEEE Access **14**, 121123-121143 (2026).
[12] S.S. Bukhari, T.M. Breuel, A. Asi and J. El-Sana, Layout analysis for Arabic historical document images using machine learning, In: Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR), 639-644 (2012).
[13] R. Saabni, A. Asi and J. El-Sana, Pattern Recognition Letters **35**, 23-33 (2014).
[14] C. Clausner, A. Antonacopoulos, N. McGregor and D. Wilson-Nunn, ICFHR 2018 competition on recognition of historical Arabic scientific manuscripts – RASM2018, In: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 471-476 (2018).

© 2026 NSP
Natural Sciences Publishing Cor.

[15] B. Kiessling, D. Stökl Ben Ezra and M.T. Miller, BADAM: a public dataset for baseline detection in Arabic-script manuscripts, In: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing (HIP), 13-18 (2019).
[16] C. Vidal-Gorène, N. Lucas, C. Salah, A. Decours-Perez and B. Dupin, RASAM – a dataset for the recognition and analysis of scripts in Arabic Maghrebi, In: ICDAR 2021 Workshops, Lecture Notes in Computer Science, Volume 12916, Springer, Cham, 265-281 (2021).
[17] C. Vidal-Gorène, C. Salah, N. Lucas, A. Decours-Perez and A. Perrier, Enhancing Arabic Maghribi handwritten text recognition with RASAM 2: a comprehensive dataset and benchmarking, In: Proceedings of the Computational Humanities Research Conference (CHR 2024), CEUR Workshop Proceedings **3834**, 200-216 (2024).
[18] A. Sulaiman, K. Omar and M.F. Nasrudin, A database for degraded Arabic historical manuscripts, In: Proceedings of the 6th International Conference on Electrical Engineering and Informatics (ICEEI), 1-6 (2017).
[19] A. Shahkolaei, A. Beghdadi, S. Al-Maadeed and M. Cheriet, MHDID: a multi-distortion historical document image database, In: Proceedings of the 2nd IEEE International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), 156-160 (2018).
[20] A. Kefali, I. Bouacha, C. Ferkous and T. Sari, Expert Systems with Applications **269**, 126453 (2025).
[21] S.A. Mahmoud, I. Ahmad, W.G. Al-Khatib, M. Alshayeb, M.T. Parvez, V. Märgner and G.A. Fink, Pattern Recognition **47**, 1096-1112 (2014).
[22] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow et al., British Medical Journal **372**, n71 (2021).
[23] K. Adam, A. Baig, S. Al-Maadeed, A. Bouridane and S. El-Menshawy, International Journal on Document Analysis and Recognition **21**, 283-290 (2018).
[24] A. Shahkolaei, H.Z. Nafchi, S. Al-Maadeed and M. Cheriet, Journal of Cultural Heritage **30**, 199-209 (2018).
[25] N. Alsaedi, B. Alrehali, H. Alahmadi and N. Abid, Historical Arabic manuscripts text recognition using convolutional neural network, In: Proceedings of the 6th Conference on Data Science and Machine Learning Applications (CDMA), 37-42 (2020).
[26] I. Ouled Omar, S. Haboubi and F. Benzarti, Toward a binarization framework resolving the Maghrebian font database challenges, In: Proceedings of the 17th International Multi-Conference on Systems, Signals & Devices (SSD), 462-466 (2020).
[27] L.S. Al-homed, K.M. Jambi and H.M. Al-Barhamtoshy, A novel dataset for known and unknown ancient Arabic manuscripts, In: Proceedings of the 20th International Conference on Language Engineering (ESOLEC), 60-65 (2022).
[28] A. Zoizou, A. Zarghili and I. Chaker, Pattern Recognition Letters **159**, 54-60 (2022).
[29] A. Droby, B. Kurar Barakat, R. Alaasam, B. Madi, I. Rabaev and J. El-Sana, Signals **3**, 535-549 (2022).
[30] B. Madi, A. Droby and J. El-Sana, International Journal on Document Analysis and Recognition **25**, 415-427 (2022).
[31] S.A. Maouloud, M.H. Ould Mohamed Dyla and C. Ba, Journal of Theoretical and Applied Information Technology **101**, 8093-8103 (2023).
[32] A. Ismail, Z. Kamel and R. Mahmoud, HICMA: the handwriting identification for calligraphy and manuscripts in Arabic dataset, In: Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023), 24-32 (2023).
[33] R. Najam and S. Faizullah, Data in Brief **56**, 110813 (2024).
[34] H. Bouchal and A. Belaid, Procedia Computer Science **244**, 57-65 (2024).
[35] M. Saeed, A. Chan, A. Mijar, J. Moukarzel, G. Habchi, C. Younes, A. Elias, C.-W. Wong and A. Khater, Muharaf: manuscripts of handwritten Arabic dataset for cursive text recognition, In: Advances in Neural Information Processing Systems **37**, Datasets and Benchmarks Track (2024).
[36] S.A. Maouloud and C. Ba, New transformer approach to the recognition of mediaeval Arabic historical manuscripts, In: Artificial Intelligence and Its Practical Applications in the Digital Economy, Lecture Notes in Networks and Systems, Volume 862, Springer, Cham, 271-283 (2024).
[37] H. Bouchal, A. Belaid and F. Meziane, ACM Transactions on Asian and Low-Resource Language Information Processing **24**, 104 (2025).
[38] A. Bouchantouf and N. Lamghari, Hierarchical segmentation of ancient Arabic manuscripts: dataset construction and deep learning evaluation, In: ICMICSA 2025, Lecture Notes in Networks and Systems, Volume 1882, Springer, Cham, 121-132 (2026).
[39] M. Naseif, A. Elhayek and A. Koubaa, MIRATH: a large and diverse Arabic manuscript dataset for text-line segmentation, In: Artificial Intelligence and its Practical Applications, Lecture Notes in Networks and Systems, Volume 1916, Springer, Cham, 334-350 (2026).
[40] J.P. Allen, J. Mullan, L. Nigst, M. Barber, T. Shahid-Khan, M. Seydi, D. Chen, Y. Weng, N. Vogler, J. Murel, O. Eshera, T. Berg-Kirkpatrick, D. Smith, S.B. Savant and M.T. Miller, Journal of Open Humanities Data **12**, 69 (2026).
[41] OpenAI, GPT-4V(ision) system card (2023), https://openai.com/research/gpt-4v-system-card.
[42] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou and J. Zhou, Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond, arXiv preprint arXiv:2308.12966 (2023).
[43] H. Liu, C. Li, Q. Wu and Y.J. Lee, Visual instruction tuning, In: Advances in Neural Information Processing Systems **36** (NeurIPS 2023) (2023).

© 2026 NSP
Natural Sciences Publishing Cor.