LIJ | *Library Insight Journal*
Vol.3 No.1 2026 *e-ISSN: 3124-8292*

Characterizing a Corpus of AI-Assisted Islamic Sermon Products: A Content Audit Toward a Library-Oriented Management Framework

Budi Nugroho
Research Center for Data and Information Science, National Research and Innovation Agency (BRIN), Bandung, Indonesia
budi045@brin.go.id

Maryulisman
Library of UIN Syarif Hidayatullah Jakarta, Indonesia
maryulisman@uinjkt.ac.id

M. Sabiq Bilhaq
Rumah Quran Program, BSI Maslahat National Zakat and Waqf Institution, Jakarta, Indonesia
sabiqbilhaq@gmail.com

Agus Miswan
Library of Politeknik Siber dan Sandi Negara, Bogor, Indonesia
agoes.miswan@poltekssn.ac.id

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Abstrak

Media keagamaan Islam berbasis kecerdasan buatan (AI) berkembang pesat namun belum terorganisasi atau diuji kualitasnya secara sistematis. Studi ini meneliti cara mengkarakterisasi, mengaudit, dan mengelola koleksi produk khotbah Islam berbasis AI dari perspektif ilmu perpustakaan dan informasi (LIS).Peneliti merancang pipeline ekstraksi fitur otomatis untuk dokumen DOCX/PDF guna menghapus kontaminasi komersial, menormalisasi nama penceramah, serta mengodekan variabel asal-usul (provenance), konten, risiko kualitas, dan etika. Dari 138 berkas koleksi praktisi, 132 berhasil diproses dan divalidasi oleh dua pakar pada 26 subsample dokumen (α Krippendorff = 1,00 untuk deskriptif; $< 0,67$ untuk risiko kualitas). Korpus didominasi oleh e-book polesan (87%) ketimbang transkrip mentah, serta terkonsentrasi pada sedikit penceramah (satu orang menyumbang 49 dari 132 dokumen). Hanya arsip keluaran (output) yang tersedia sehingga metrik proses hulu tidak dapat dipulihkan (provenance gap). Atribusi sumber tidak merata, dan 79% produk (104 dari 132) memuat iklan komersial tersemat.Arsip konten da'wah berbasis AI yang hanya berupa keluaran merupakan sumber informasi yang dapat dikelola, namun kurang terdeskripsi dan tidak seimbang. Menggabungkan Model Siklus Hidup Kurasi DCC, metadata OAIS/PREMIS, serta teori kualitas informasi, studi ini mengusulkan kerangka kerja LIS yang mencakup penangkapan metadata, pencatatan asal-usul, daftar periksa validasi, serta pelacakan etika dan izin guna mengatasi risiko kurasi konten bermediasi AI.

Kata Kunci: Manajemen konten digital, kurasi digital, analisis korpus, metadata, provenance, kualitas informasi, media keagamaan berbasis AI, da'wah digital.

Abstract

Islamic religious media produced with the assistance of artificial intelligence (AI) is growing rapidly, yet it is rarely organized or systematically evaluated for quality. This study investigates how to characterize, audit, and manage a collection of AI-assisted Islamic sermon products from a library and information science (LIS) perspective. The researchers designed an automated feature-extraction pipeline for DOCX/PDF documents to remove commercial contamination, normalize speaker names, and encode variables related to provenance, content, quality risk, and ethics. Out of 138 practitioner files, 132 were successfully processed and validated by two experts on a 26-document subsample (Krippendorff's α = 1.00 for descriptive variables; < 0.67 for quality-risk judgments). The corpus is dominated by polished e-books (87%) rather than raw transcripts, and is heavily skewed toward a few speakers (one speaker accounts for 49 out of 132 documents). Only output archives are available, meaning upstream process metrics cannot be recovered (creating a provenance gap). Source attribution is uneven, and 79% of the products (104 out of 132) contain embedded commercial advertisements. An output-only archive of AI-based da'wah content is a manageable but under-described and imbalanced information resource. Combining the DCC Curation Lifecycle Model, OAIS/PREMIS metadata, and information-quality theory, this study proposes an LIS framework encompassing metadata capture, provenance recording, a validation checklist, and ethics and permission tracking to address the curation risks of AI-mediated content.

Keywords: Digital content management, digital curation, corpus analysis, metadata, provenance, information quality, AI-assisted religious media, digital da'wah.

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

INTRODUCTION

Generative artificial intelligence (AI) has lowered the cost of turning spoken religious lectures into reusable digital products. Preachers, volunteers, and small da'wah teams now routinely convert recorded sermons into e-books, summaries, slides, and social-media materials with the help of automatic speech recognition (ASR) and large language models (LLMs). For libraries, archives, and community information units, this creates a new and largely unmanaged category of born-digital material: heterogeneous, rapidly growing, and produced outside any cataloging or quality-control regime. The central question of this study is how such a collection can be described, audited, and governed as a digital knowledge resource.

The motivation is concrete. A practitioner on our team accumulated 138 sermon-related files, stored in nested, speaker-named folders as DOCX and PDF, over roughly one year. Inspection revealed that the archive preserved only finished outputs; audio recordings, intermediate transcripts, processing times, and ASR configurations had not been retained. This makes efficiency-style evaluation impossible but raises questions squarely within library and information science (LIS): what is actually in the collection, how complete is its metadata, how sound is its scriptural sourcing, and what ethical and provenance risks does it carry?

Existing literature supplies partial building blocks but not an answer. Studies on AI and natural language processing for Islamic knowledge demonstrate the feasibility of computational analysis of Islamic textual resources (Abdhood et al., 2025; Al-Qerem et al., 2024; Athallah & Lhaksmana, 2025; Badry et al., 2021; Mohamed et al., 2022). Work on ASR and text summarization documents the accuracy and factual-drift risks that make downstream validation necessary (Ahlawat et al., 2025; Altameemi et al., 2025; Kheddar et al., 2024; Kumar & Solanki, 2023; Radford et al., 2023; Supriyono et al., 2024). Research on retrieval-augmented generation and agentic AI shows how grounding and orchestration can improve reliability (Huang & Huang, 2025; Lewis et al., 2020; Plaat et al., 2025; Wang et al., 2024; Xi et al., 2025). Within LIS, scholarship on AI has concentrated on how libraries adopt intelligent technologies in their services and operations (Cox et al., 2019; Cox & Mazumdar, 2024), whereas the arrival of machine-generated material in collections has only recently been documented as a collection-development problem (Chau, 2025). What remains under-examined is the information-management side: once such products exist in large numbers, how are they described, quality-assured, and governed?

This study approaches that question through established LIS frameworks rather than through computational performance metrics. Digital curation

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

provides the organizing lens: the DCC Curation Lifecycle Model treats curation as a sequence of actions from conceptualization and creation through appraisal, ingest, preservation, and reuse, with description and preservation planning running across all stages (Higgins, 2008; Yakel, 2007). The Open Archival Information System (OAIS) reference model and the PREMIS data dictionary specify the provenance, context, and rights information an object must carry to remain understandable and trustworthy over time (CCSDS, 2012; PREMIS Editorial Committee, 2015), and the W3C PROV model expresses provenance as relations among entities, activities, and agents (Moreau & Missier, 2013). Records continuum thinking adds that evidential value depends on capture at the moment of creation rather than on custody after the fact (Upward, 1996), a point echoed in work on the contextual information that digital collections require (Lee, 2011). For content quality, we draw on the multidimensional view of information quality, spanning intrinsic, contextual, representational, and accessibility dimensions (Wang & Strong, 1996), and on the notion of cognitive authority, through which users judge whom and what to trust (Rieh, 2002). Finally, professional information ethics frames obligations of integrity, transparency, and respect for creators' rights (IFLA, 2012), while the FAIR principles articulate what makes a digital resource findable and reusable (Wilkinson et al., 2016). Together, these lenses define what a well-managed collection of AI-assisted religious products should look like, and therefore what our audit measures.

The novelty of this study lies in its object and its lens. To our knowledge, it is the first study to treat an organically accumulated, output-only collection of AI-assisted religious products as an object of LIS inquiry: prior work has examined either the computational processing of Islamic texts or the adoption of AI within library services, but not the description, quality assurance, and governance of AI-assisted religious outputs once they become a collection. Within this framing, the study makes three empirical and methodological contributions. First, it designs and applies a reproducible pipeline that extracts management-relevant features from AI-assisted da'wah products in their native DOCX and PDF formats, including automatic removal of commercial contamination. Second, it characterizes a real 138-file corpus—its composition, speaker distribution, metadata recoverability, and content properties. Third, it surfaces the quality, provenance, and ethics risks that an LIS-oriented management framework must address, and proposes that framework. Its theoretical contribution is to extend digital curation and provenance theory to AI-mediated content, identifying the output-only provenance gap as a distinctive curation failure and relocating the library's curatorial role upstream, to the point of creation. In doing so, it reframes AI-assisted sermon processing from a question of production efficiency into one of digital collection management, where the library acts as curator and quality gatekeeper.

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

METHOD

Research Design and Corpus

We adopt a design-oriented, practice-based case study. The design artifact is a feature-extraction and coding pipeline; the object of study is a practitioner-managed collection of AI-assisted Islamic sermon products accumulated since late 2025 and stored as DOCX and PDF files in speaker-named subfolders. The collection comprised 138 files, of which 132 were successfully processed; six files could not be extracted because their content is image-based (scanned) and are excluded from the statistics below pending optical character recognition (OCR).

Unit of Analysis and Coding Scheme

The unit of analysis is a single finished da'wah product (one document). Each document is coded across 36 variables in four families. Provenance variables capture identity, format, source program, speaker name(s) and type, estimated date and its source, and any stated AI involvement. Content and structure variables record theme and thematic domain, word count, table-of-contents and question-and-answer presence, chapter count, and a processing-state label (raw ASR, lightly edited, compiled from text, or polished e-book). Islamic-content variables record the presence and density of Arabic passages and Qur'anic and hadith citations, with an attribution-quality rating. Quality-risk and ethics variables flag probable Qur'anic attribution errors, corrupted Arabic rendering, unverified claims, sensitive topics, embedded commercial content, and a derived validation-priority level.

Feature-Extraction Pipeline

The pipeline is implemented in Python and reads both DOCX and PDF. A key preprocessing step removes commercial contamination: in this collection, a gold-selling advertisement had been appended to the end of many products. The pipeline detects this block using stable text markers and separates it from the sermon content before any counting, logging the removed character count for transparency. It then computes text-derived features (word count, chapter and Q&A detection, Arabic-character ratio, keyword-based Qur'an and hadith citation counts, and date estimation), applies heuristic speaker and theme detection, and normalizes speaker names so that spelling and title variants (for example, "KH. Bachtiar Nasir" and "Ustadz Bachtiar Nasir") collapse to a single canonical form. All counts operate on the cleaned text, so commercial noise does not distort them.

Validation of the Automated Coding

Because automated coding can err, the hybrid and quality-risk variables were validated by expert review using a formal coding protocol and codebook.

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Two reviewers independently re-coded a random subsample of 26 documents (20% of the valid corpus), blind to the automatic codes, confirming or correcting the theme, speaker type, processing state, and attribution-quality fields and adjudicating each quality-risk flag as confirmed or rejected. Agreement was assessed with Krippendorff's α, using a nominal metric for the categorical variables and an ordinal metric for the five-point meaning-faithfulness rating, and interpreted against the conventional thresholds of 0.80 for reliable and 0.667 for tentative conclusions (Krippendorff, 2018). Agreement was perfect ($\alpha = 1.00$) for all four descriptive variables—thematic domain, speaker type, processing state, and attribution quality—but fell below the tentative threshold for every quality-risk judgment: probable Qur'anic attribution error ($\alpha = 0.17$), corrupted Arabic rendering (0.32), unverified claim (−0.06), sensitive topic (0.27), and meaning faithfulness (−0.43). Disagreements were reconciled through discussion and the ambiguous rules refined; the quality-risk results below should nonetheless be read with caution. The figures reported below reflect the validated coding. Consistent with the protocol, automatic flags are treated as triage signals that direct expert attention rather than as confirmed findings.

RESULTS AND DISCUSSION

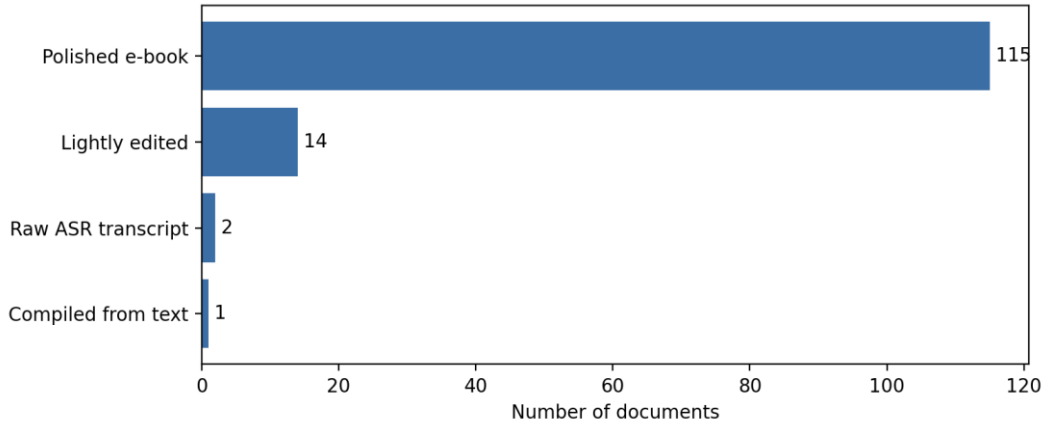
Of 138 files in the collection, 132 (95.7%) were processed successfully; six could not be extracted because their text is image-based and are excluded from the figures below. The pipeline produced a complete record per document, and expert review confirmed the descriptive coding with perfect agreement, whereas the quality-risk fields showed low inter-coder agreement (see Method) and are interpreted with caution. The results are organized around four themes: corpus composition, speaker concentration, metadata and provenance, and content quality and ethics.

Corpus Composition

The corpus is overwhelmingly composed of finished products rather than raw material. As Figure 1 shows, 115 of 132 documents (87%) are polished e-books—complete with cover pages, tables of contents, and numbered chapters—while only 14 are lightly edited notes, two are raw ASR transcripts, and one is compiled from a fiqh reference text rather than a spoken sermon. Formats are predominantly PDF (107) over DOCX (25). This distribution has a direct management implication: the archive is a collection of publication-grade outputs whose underlying process is no longer visible, so any quality assurance must be performed retrospectively on the products themselves rather than during production.

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Figure 1. Distribution of documents by processing state (n = 132)



Source: Authors' pipeline output (2026)

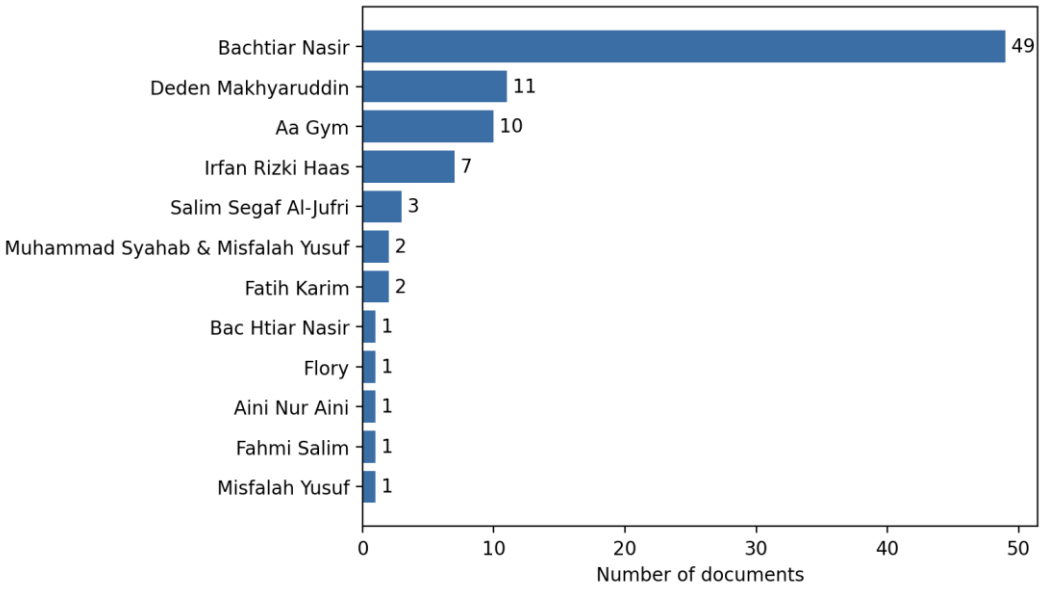
Thematically the collection concentrates on a few areas: worship and fiqh account for 29% of documents, character and tazkiyah for 25%, and tafsir or muhasabah reflection for 17%, with smaller shares for community, motivation, health, and socio-political content. The average document contains 903 words (median 779, range 216–5,546), and 81% include a table of contents—further evidence that the products are structured, book-like artifacts rather than raw transcripts.

Speaker Concentration

The most consequential structural finding is the concentration of the corpus around very few speakers. After name normalization, only sixteen distinct speakers appear across 132 documents, and a single preacher accounts for 49 of them—more than a third of the entire collection—followed by eleven and ten documents for the next two speakers (Figure 2). A speaker could be identified for 93 documents (70%); the remainder, including the raw transcripts and the fiqh compilation, name no speaker. This imbalance matters for collection management and for any downstream analytical use: a corpus dominated by one voice is not a neutral sample of Indonesian da'wah, and metadata describing speaker coverage is essential context for anyone who later reuses the collection.

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Figure 2. Document count by speaker after name normalization (top speakers)



Source: Authors' pipeline output (2026)

Metadata Recoverability and the Provenance Gap

Metadata recovery was uneven. A usable date was obtained for 93 of 132 documents (70%), most often parsed from in-text content and only occasionally from filename prefixes; the remainder yielded no date, and in several products the only date present belonged to the embedded advertisement rather than to the sermon—a trap the pipeline was designed to avoid. Crucially, none of the upstream process attributes that a production-focused study would rely on—audio duration, manual and AI processing time, ASR tool and accuracy—could be recovered from any document, because the archive preserved only outputs. From an LIS standpoint this systematic provenance gap limits accountability and reuse. The practical lesson is that provenance metadata must be captured at creation time; it cannot be reconstructed reliably after the fact.

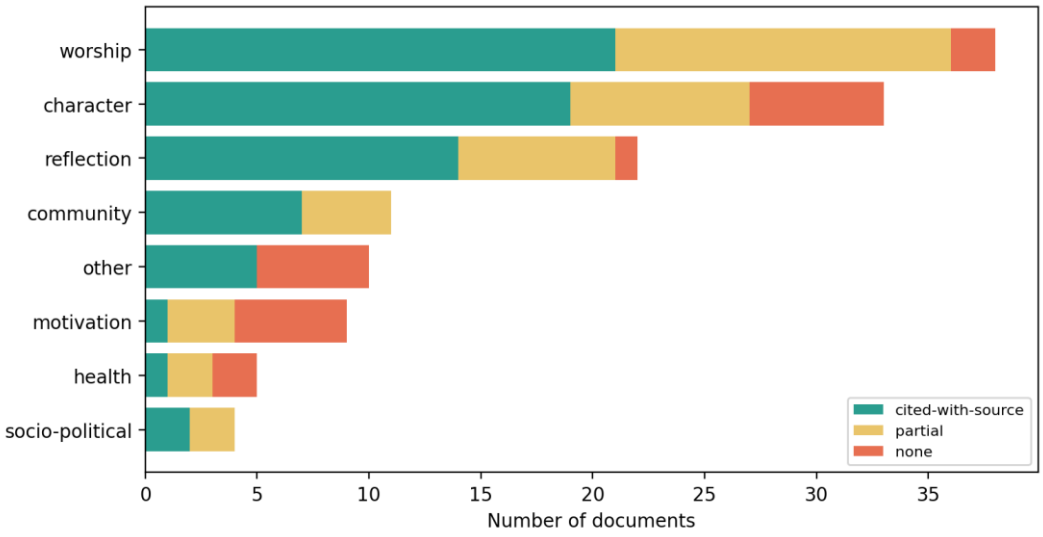
Content Quality

Scriptural sourcing is present but uneven. Of 132 documents, 70 (53%) cite named hadith transmitters or specific verse references, 41 (31%) reference verses and hadith only generically, and 21 (16%) provide no scriptural source at all. Figure 3 breaks this down by thematic domain and shows that sourcing is strongest in worship, character, and reflection content—where citation is expected—and weakest in motivation and health material, precisely the domains where unsupported claims are most likely to mislead. The pipeline detected 464 Qur'anic and 436 hadith references overall, and expert review identified isolated accuracy problems, including a probable Qur'anic mislabel (a verse on purposeful creation attributed to the wrong surah) and a cluster of unverified health claims in a document on cupping therapy. Corrupted Arabic rendering

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

also appears in a number of PDFs where devotional Arabic was expected but lost to encoding; this remains a data-quality caveat requiring case-by-case confirmation.

Figure 3. Attribution quality by thematic domain (n = 132)



Source: Authors' pipeline output (2026)

Table 1 consolidates the corpus profile, and Table 2 summarizes the attribution and validation-priority distributions that structure the retrospective audit. The validation-priority score, derived from the confirmed quality-risk flags, places two documents at high priority, 70 at medium, and 60 at low, giving reviewers a workable queue rather than an undifferentiated mass of 132 files. Because the underlying quality-risk judgments showed low inter-coder agreement, this distribution should be read as a provisional review queue rather than as a measured risk profile.

Table 1. Profile of the AI-assisted da'wah corpus (valid n = 132 of 138)

Attribute	Value
Files collected / successfully processed	138 / 132 (6 image-based excluded)
Format	PDF 107, DOCX 25
Processing state	polished e-book 115, lightly edited 14, raw ASR 2, compiled-from-text 1
Distinct speakers (normalized)	16; top speaker 49 documents (37%)
Speaker identifiable	93 of 132 (70%)
Leading thematic domains	worship/fiqh 29%, akhlak 25%, reflection 17%
Mean / median word count	903 / 779 words (range 216–5,546)
Documents with table of contents	107 of 132 (81%)
Arabic passages present	69 of 132 (52%)

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Qur'an / hadith citations detected	464 / 436
Commercial advertisement embedded	104 of 132 (79%)

Source: Authors' pipeline output (2026)

Table 2. Attribution quality and validation-priority distribution (n = 132)

Item	Count (share)
Attribution: cited-with-source	70 (53%)
Attribution: partial / generic	41 (31%)
Attribution: none	21 (16%)
Validation priority: high	2 (2%)
Validation priority: medium	70 (53%)
Validation priority: low	60 (45%)
Sensitive-topic documents	10 (8%)
AI involvement disclosed	9 (7%)

Source: Authors' pipeline output (2026)

Ethics, Commercialization, and the Library Frame

The most pervasive ethics finding is embedded commercialization: 104 of 132 products (79%) carried an appended gold-selling advertisement, from which the pipeline removed roughly 105,000 characters of commercial text in total. Read through the information-quality lens, blending unrelated promotion into religious learning material degrades the believability and reputation dimensions of quality (Wang & Strong, 1996) and, as noted, also introduces metadata error. It further disturbs the cues on which users base cognitive-authority judgments—speaker identity, scriptural sourcing, and the absence of vested interest (Rieh, 2002)—which in religious content are central to whether material is trusted. Beyond advertising, ten documents (8%) addressed sensitive socio-political subject matter requiring careful dissemination, and AI involvement was explicitly disclosed in only nine (7%), meaning that for most products a reader is not told the content was AI-assisted. From the standpoint of professional information ethics, which commits libraries to integrity, transparency about sources, and respect for creators' rights (IFLA, 2012), undisclosed AI mediation and unrecorded permission status are not peripheral omissions but curation defects; a parallel case has already been reported in an academic library, where AI-generated derivative e-books entered the catalog unrecognized (Chau, 2025). These observations show why an information-management framework must treat permission status, source attribution, commercial-content separation, AI disclosure, and sensitive-topic handling as first-class descriptive fields. Each maps onto an established library function—descriptive and administrative metadata, provenance and records management, rights management, and

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

collection quality control—positioning the library as curator and quality gatekeeper for AI-assisted religious learning collections.

Toward a Management Framework

Synthesizing the findings, we propose a lightweight LIS-oriented framework with four components, each anchored in an established LIS framework rather than devised ad hoc (Table 3). First, metadata capture at creation: speaker, date, source program, and permission recorded when a product is made, using a descriptive schema such as Dublin Core (DCMI Usage Board, 2020). This corresponds to the create-or-receive action of the DCC Curation Lifecycle Model (Higgins, 2008) and to the records-continuum principle that evidence must be captured when it is created (Upward, 1996), and it directly closes the provenance gap documented above. Second, provenance recording that links each product to its source recording and to each processing step (ASR, LLM summarization, and human editing), so that quality can be traced. In OAIS terms this is Preservation Description Information (CCSDS, 2012); in PREMIS and W3C PROV terms, each AI or human step is an event or activity associated with an agent (Moreau & Missier, 2013; PREMIS Editorial Committee, 2015). Third, a validation checklist—the audit instrument used here—covering Arabic terms, Qur'anic and hadith accuracy, speaker attribution, meaning preservation, and ethical compliance; it operationalizes the appraise-and-select action of the curation lifecycle and the intrinsic and representational dimensions of information quality (Wang & Strong, 1996). Fourth, ethics and permission tracking, including explicit separation of any commercial content and disclosure of AI involvement, recorded as rights and agent metadata (PREMIS Editorial Committee, 2015) and grounded in professional information ethics (IFLA, 2012). Applied prospectively, this framework would make the collection findable and reusable in the sense of the FAIR principles (Wilkinson et al., 2016), converting an opaque folder of outputs into a governed, reusable digital collection.

Table 3. Mapping of the proposed management framework to established LIS frameworks

Component	Corpus evidence addressed	Anchoring LIS framework
1. Metadata capture at creation	Date recoverable for 70% and speaker for 70% of documents; no process metadata in any document	DCC lifecycle "create or receive" (Higgins, 2008); records continuum (Upward, 1996); Dublin Core (DCMI Usage Board, 2020)
2. Provenance recording	Upstream process attributes (audio, ASR tool, processing time) unrecoverable for all 132 documents	OAIS Preservation Description Information (CCSDS, 2012); PREMIS events and agents (PREMIS Editorial Committee, 2015); W3C PROV (Moreau & Missier, 2013)

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Component	Corpus evidence addressed	Anchoring LIS framework
3. Validation checklist	16% without scriptural source, 31% generic only; mislabelled verse; unverified health claims	DCC "appraise and select" (Higgins, 2008); information-quality dimensions (Wang & Strong, 1996); cognitive authority (Rieh, 2002)
4. Ethics and permission tracking	Advertisement embedded in 79%; AI involvement disclosed in 7%; sensitive topics in 8%	PREMIS rights (PREMIS Editorial Committee, 2015); IFLA Code of Ethics (IFLA, 2012); FAIR reuse conditions (Wilkinson et al., 2016)

Source: Authors' synthesis (2026)

Theoretical Contribution

Beyond its practical framework, the study makes three contributions to LIS theory. First, it extends the digital curation lifecycle to AI-mediated content. The DCC model treats conceptualization and creation as curation actions (Higgins, 2008), but AI-assisted production compresses recording, transcription, summarization, and design into a rapid, tool-driven sequence that leaves no trace once the output is saved. We term the resulting condition an output-only provenance gap: a collection whose objects are complete as products but evidentially empty as records, because every intermediate entity and activity that PROV or PREMIS would describe has been discarded (Moreau & Missier, 2013; PREMIS Editorial Committee, 2015). Unlike the incidental gaps of legacy collections, this one is generated systematically by the production technology itself, which implies that AI processing steps should be treated as preservation-relevant events to be captured by default.

Second, the study connects information-quality theory with religious information. Attribution quality in da'wah material functions as an operational form of cognitive authority (Rieh, 2002): verse and hadith references, a named speaker, and the absence of commercial interest are the cues readers use to decide whether religious content is trustworthy. Our finding that sourcing is weakest in motivation and health material, together with near-universal commercial embedding and rare AI disclosure, shows how AI-assisted production can erode exactly these cues while raising the representational polish of the product. Intrinsic and representational quality, which Wang and Strong (1996) treat as complementary dimensions, here move in opposite directions.

Third, the study supports a post-custodial repositioning of the library in this domain. In records-continuum terms, governance must operate at the create and capture dimensions rather than only at the point of custody (Upward, 1996; Lee, 2011). For AI-assisted religious media, this means the library's curatorial role begins upstream, advising producers on metadata, provenance, and disclosure

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

at creation, rather than with retrospective cleanup of finished outputs. Methodologically, the 36-variable coding scheme and its codebook offer a transferable audit instrument for other collections of AI-assisted content.

Limitations

Several limitations qualify these findings. Six image-based documents await OCR and are excluded here. Automated theme, speaker, and quality coding, although validated on a subsample by expert review, is not error-free, and inter-coder agreement on the quality-risk judgments was low (α from −0.43 to 0.32), which may partly reflect the rarity of positive cases in a 26-document subsample but also signals that these judgments need sharper decision rules and coder calibration before they can serve as prevalence measures; keyword-based citation counting can miss or over-count references, and the corrupted-Arabic heuristic over-flags and requires case-by-case confirmation. The corpus is a single practitioner's collection and is not representative of Indonesian da'wah more broadly, as its speaker concentration makes clear. Finally, the study characterizes and audits the collection but does not evaluate the accuracy of the underlying AI tools, which would require paired reference data that the archive does not contain.

CONCLUSION

This study characterized and audited a real collection of AI-assisted Islamic sermon products preserved as outputs only. Using a reproducible extraction and coding pipeline, validated by expert review and applied to 132 of 138 documents, we found that the corpus is dominated by finished e-books, heavily concentrated on a small number of speakers, and systematically missing upstream provenance; that scriptural attribution is uneven and weakest where unsupported claims are most consequential; and that commercial advertising is embedded in four of five products. These findings support an LIS-oriented management framework built on metadata capture, provenance recording, a validation checklist, and ethics and permission tracking. Anchored in the DCC Curation Lifecycle Model, OAIS and PREMIS provenance metadata, and information-quality theory, the study contributes the concept of an output-only provenance gap and argues for relocating library curation upstream, to the point at which AI-assisted products are created. More broadly, the study shows that the rapid growth of AI-assisted religious media is not only a production phenomenon but an information-management challenge, and that libraries are well placed to meet it as curators and quality gatekeepers. Future work will complete OCR for the remaining documents, extend the audit to comparable collections, and pilot the proposed framework in a working da'wah content service.

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

REFERENCES

Abdhood, S. F., et al. (2025). Data augmentation for Arabic text classification: A review of current methods, challenges and prospective directions. PeerJ Computer Science. https://doi.org/10.7717/peerj-cs.2685

Ahlawat, H., Aggarwal, N., & Gupta, D. (2025). Automatic speech recognition: A survey of deep learning techniques and approaches. International Journal of Cognitive Computing in Engineering, 6, 201–237. https://doi.org/10.1016/j.ijcce.2024.12.007

Al-Qerem, A., Raja, M., Taqatqa, S., & Sara, M. R. A. (2024). Utilizing deep learning models (RNN, LSTM, CNN-LSTM, and Bi-LSTM) for Arabic text classification. In Advances in Intelligent Systems and Computing (pp. 287–301). Springer. https://doi.org/10.1007/978-3-031-43490-7_22

Altameemi, Y., Altamimi, M., Alkhalil, A., Uliyan, D., & Mansour, R. F. (2025). Text summarization method of argumentative discourse by combining the BERT-transformer model. Frontiers in Artificial Intelligence, 8, 1654496. https://doi.org/10.3389/frai.2025.1654496

Athallah, M. R., & Lhaksmana, K. M. (2025). Hadith text classification based on topic using convolutional neural network (CNN) and TF-IDF. Journal of Renewable Energy, Electrical, and Computer Engineering, 5(1). https://doi.org/10.29103/jreece.v5i1.20354

Badry, R. I., Alshareef, A. M., Fattah, A. A. A., & Almazroi, A. A. (2021). A systematic review on hadith authentication and classification methods. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3434236

CCSDS. (2012). Reference model for an Open Archival Information System (OAIS) (CCSDS 650.0-M-2, Magenta Book). Consultative Committee for Space Data Systems. https://public.ccsds.org/Pubs/650x0m2.pdf

Chau, S. (2025). Unexpected AI in the collections: A collection development case study on navigating a new academic information ecosystem. IFLA World Library and Information Congress. https://repository.ifla.org/handle/20.500.14598/6860

Cox, A. M., & Mazumdar, S. (2024). Defining artificial intelligence for librarians. Journal of Librarianship and Information Science, 56(2), 330–340. https://doi.org/10.1177/09610006221142029

Cox, A. M., Pinfield, S., & Rutter, S. (2019). The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries. Library Hi Tech, 37(3), 418–435. https://doi.org/10.1108/LHT-08-2018-0105

DCMI Usage Board. (2020). DCMI metadata terms. Dublin Core Metadata Initiative. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Higgins, S. (2008). The DCC Curation Lifecycle Model. International Journal of Digital Curation, 3(1), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

Huang, Y., & Huang, J. X. (2025). A survey on retrieval-augmented text generation for large language models. ACM Computing Surveys. https://doi.org/10.1145/3805774

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

IFLA. (2012). IFLA code of ethics for librarians and other information workers (full version). International Federation of Library Associations and Institutions. https://repository.ifla.org/handle/20.500.14598/1850

Kheddar, H., Hemis, M., & Himeur, Y. (2024). Automatic speech recognition using advanced deep learning approaches: A survey. Information Fusion, 109, 102422. https://doi.org/10.1016/j.inffus.2024.102422

Krippendorff, K. (2018). Content analysis: An introduction to its methodology (4th ed.). SAGE Publications.

Kumar, S., & Solanki, A. (2023). An abstractive text summarization technique using transformer model with self-attention mechanism. Neural Computing and Applications, 35(25), 18603–18622. https://doi.org/10.1007/s00521-023-08687-7

Lee, C. A. (2011). A framework for contextual information in digital collections. Journal of Documentation, 67(1), 95–143. https://doi.org/10.1108/00220411111105470

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474). https://doi.org/10.48550/arXiv.2005.11401

Mohamed, E., Ab Aziz, M., & Mohamad, M. (2022). Linguistic features evaluation for hadith authenticity through automatic machine learning. Digital Scholarship in the Humanities, 37(3), 830–843. https://doi.org/10.1093/llc/fqab086

Moreau, L., & Missier, P. (Eds.). (2013). PROV-DM: The PROV data model (W3C Recommendation). World Wide Web Consortium. https://www.w3.org/TR/prov-dm/

Plaat, A., van Duijn, M., van Stein, N., Preuss, M., van der Putten, P., & Batenburg, K. J. (2025). Agentic large language models: A survey. Journal of Artificial Intelligence Research, 84. https://doi.org/10.1613/jair.1.18675

PREMIS Editorial Committee. (2015). PREMIS data dictionary for preservation metadata (Version 3.0). Library of Congress. https://www.loc.gov/standards/premis/v3/

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 28492–28518). PMLR. https://doi.org/10.48550/arXiv.2212.04356

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. Journal of the American Society for Information Science and Technology, 53(2), 145–161. https://doi.org/10.1002/asi.10017

Supriyono, Wibawa, A. P., Suyono, & Kurniawan, F. (2024). A survey of text summarization: Techniques, evaluation and challenges. Natural Language Processing Journal, 7, 100070. https://doi.org/10.1016/j.nlp.2024.100070

Upward, F. (1996). Structuring the records continuum, part one: Post-custodial principles and properties. Archives and Manuscripts, 24(2), 268–285.

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. (2025). The rise and potential of large language model based agents: A survey. Science China Information Sciences. https://doi.org/10.1007/s11432-023-3911-2

Yakel, E. (2007). Digital curation. OCLC Systems & Services, 23(4), 335–340. https://doi.org/10.1108/10650750710831466

DOI: https://doi.org/10.15408/lij.
© 2024 The Author(s). Published by (Institution). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)