Ameera Yahya Ahmed AL-Hooti[1]

THE INTELLIGENT ARCHIVAL ECOSYSTEM: DESIGNING THE FUTURE ARCHITECTURE OF AI-POWERED ARCHIVES

Abstract

Purpose: *This article reviews how artificial intelligence (AI) technologies are reshaping the theory and practice of archival science and, on the basis of that review, proposes a conceptual reference architecture for what is termed an "intelligent archival ecosystem" – an integrated model in which AI-based tools support, rather than replace, the professional judgement of archivists across the record lifecycle.*

Method/approach: *The study combines three components: (1) a structured narrative review of scholarly, professional and policy literature on AI in archives, records management and digital preservation, drawing mainly on sources published between 2019 and 2026, including prior systematic reviews, practitioner studies, standards and official guidance from national archives, together with a cross-study analytical synthesis of eleven empirical evaluations; (2) a documentary analysis of six publicly reported implementations and policy responses, used to test and refine the architecture; and (3) design reasoning grounded in archival theory, including the records continuum model, the OAIS reference model (ISO 14721:2025), Records in Contexts and ISO 15489.*

Results: *The review shows that natural language processing and computer vision are the most mature techniques; that accuracy depends strongly on the interpretive demands of the task (in one national-library experiment, up to 99% for title and author fields but below 50% for subject and genre); and that the provenance of AI-generated metadata and governance mechanisms are rarely evaluated in published implementations. The article proposes a five-layer architecture – ingestion and capture, AI-augmented description and appraisal, trust and provenance, discovery and access, and governance – that situates natural language processing, computer vision, machine-learning classification and large language models within a human-in-the-loop framework accountable to the archival principles of authenticity, reliability, integrity and usability, and it*

[1] Ameera Yahya Ahmed AL-Hooti, PhD student at Alma Mater Europaea University, email: ameera1yahya2@gmail.com.

introduces field-level automation tiers as a design principle. Experimental and interview evidence indicates that recognition and description quality depend on adaptation to the script and the collection, and that hands-on exposure to imperfect AI output lowers participants' trust and willingness to use AI. A protocol for empirical validation through multiple case studies is specified.

Conclusions/findings: *AI can substantially increase the efficiency and accessibility of archival description, appraisal and retrieval, but its responsible integration requires explicit governance structures, transparency mechanisms, the treatment of AI-related materials as records in their own right, and a redefined professional role for the archivist as orchestrator and auditor of automated processes rather than as sole executor of archival functions. Because the model is conceptual and evidence-informed rather than empirically validated, the article identifies multi-institutional case-study research as the priority for future work.*

Keywords: *artificial intelligence, archival science, records management, digital archives, archival description, human-in-the-loop, trustworthy AI*

INTELIGENTNI ARHIVSKI EKOSISTEM: NAČRTOVANJE PRIHODNJE ARHITEKTURE ARHIVOV, PODPRTIH Z UMETNO INTELIGENCO

Izvleček

Namen: *Članek pregleduje, kako tehnologije umetne inteligence (UI) spreminjajo teorijo in prakso arhivske znanosti, ter na podlagi tega pregleda predlaga konceptualno referenčno arhitekturo t. i. »inteligentnega arhivskega ekosistema«. Gre za celovit model, v katerem orodja, ki temeljijo na umetni inteligenci, podpirajo – in ne nadomeščajo – strokovne presoje arhivistov skozi celoten življenjski cikel dokumentov.*

Metoda/pristop: *Raziskava združuje tri sestavine: (I) strukturiran narativni pregled znanstvene, strokovne in politične literature o uporabi umetne inteligence v arhivih, pri upravljanju dokumentov in digitalnem ohranjanju, ki temelji predvsem na virih, objavljenih med letoma 2019 in 2026, vključno s prejšnjimi sistematičnimi pregledi, raziskavami med strokovnjaki, standardi in uradnimi smernicami nacionalnih arhivov, skupaj z medštudijsko analitično sintezo enaj-*

stih empiričnih vrednotenj; (2) dokumentarno analizo šestih javno predstavljenih izvedb in političnih odzivov, s katero se arhitektura preverja in dopolnjuje; ter (3) konceptualno načrtovanje, ki izhaja iz uveljavljene arhivske teorije, vključno z modelom records continuum, referenčnim modelom OAIS (ISO 14721:2025), standardom Records in Contexts in standardom ISO 15489.

Rezultati: *Pregled kaže, da sta obdelava naravnega jezika in računalniški vid najbolj zreli tehniki, da je natančnost močno odvisna od interpretativne zahtevnosti naloge (v enem poskusu nacionalne knjižnice do 99 % pri poljih naslova in avtorja, a manj kot 50 % pri poljih predmeta in zvrsti) ter da se provenienca metapodatkov, ustvarjenih z umetno inteligenco, in mehanizmi upravljanja v objavljenih izvedbah redko vrednotijo. Članek predstavlja petplastno arhitekturo, ki vključuje sprejem in zajem, opisovanje in vrednotenje, podprto z umetno inteligenco, zaupanje in provenienco, iskanje in dostop ter upravljanje. V to arhitekturo umešča obdelavo naravnega jezika, računalniški vid, klasifikacijo s pomočjo strojnega učenja in velike jezikovne modele, pri čemer ohranja človeka v procesu odločanja ter zagotavlja skladnost z arhivskimi načeli avtentičnosti, zanesljivosti, celovitosti in uporabnosti. Kot načrtovalsko načelo uvaja ravni avtomatizacije po posameznih poljih. Eksperimentalni in intervjujski dokazi kažejo, da sta kakovost razpoznavanja in opisovanja odvisni od prilagoditve pisavi in zbirki ter da neposredna izkušnja z nepopolnimi izhodi umetne inteligence zmanjša zaupanje udeležencev in njihovo pripravljenost za uporabo. Določen je tudi protokol za empirično preverjanje z več študijami primera.*

Sklepi/ugotovitve: *Umetna inteligenca lahko bistveno poveča učinkovitost in dostopnost arhivskega popisovanja, vrednotenja in iskanja. Vendar njeno odgovorno vključevanje zahteva jasno določene strukture upravljanja, mehanizme preglednosti, obravnavo gradiva, povezanega z umetno inteligenco, kot dokumentarnega gradiva samega po sebi, ter na novo opredeljeno poklicno vlogo arhivista, ki postaja koordinator in presojevalec avtomatiziranih postopkov, ne pa njihov edini izvajalec. Ker je model konceptualen in temelji na dokazih, ni pa empirično preverjen, članek kot prednostno usmeritev prihodnjih raziskav določa večinstitucionalne študije primera.*

Ključne besede: *umetna inteligenca, arhivska znanost, upravljanje dokumentov, digitalni arhivi, arhivsko popisovanje, človek v zanki, zaupanja vredna umetna inteligenca*

L'ECOSISTEMA ARCHIVISTICO INTELLIGENTE: PROGETTARE L'ARCHITETTURA FUTURA DEGLI ARCHIVI BASATI SULL'INTELLIGENZA ARTIFICIALE

Obiettivo: *Il presente articolo passa in rassegna il modo in cui le tecnologie di intelligenza artificiale (IA) stanno ridefinendo la teoria e la pratica della scienza archivistica e, sulla base di tale rassegna, propone un'architettura di riferimento concettuale per quello che viene definito un "ecosistema archivistico intelligente" – un modello integrato in cui gli strumenti basati sull'IA supportano, anziché sostituire, il giudizio professionale degli archivisti lungo l'intero ciclo di vita dei documenti.*

Metodo/approccio: *Lo studio combina tre componenti: (1) una rassegna narrativa strutturata della letteratura scientifica, professionale e di policy sull'IA negli archivi, nella gestione dei documenti e nella conservazione digitale, basata principalmente su fonti pubblicate tra il 2019 e il 2026, comprese precedenti rassegne sistematiche, studi sui professionisti, standard e linee guida ufficiali di archivi nazionali, insieme a una sintesi analitica trasversale di undici valutazioni empiriche; (2) un'analisi documentale di sei implementazioni e risposte di policy rese pubbliche, utilizzata per verificare e affinare l'architettura; e (3) un ragionamento di progettazione fondato sulla teoria archivistica consolidata, tra cui il modello del continuum dei documenti, il modello di riferimento OAIS (ISO 14721:2025), Records in Contexts e ISO 15489.*

Risultati: *La rassegna mostra che l'elaborazione del linguaggio naturale e la visione artificiale sono le tecniche più mature; che l'accuratezza dipende fortemente dalla complessità interpretativa del compito (in un esperimento di una biblioteca nazionale, fino al 99% per i campi titolo e autore ma inferiore al 50% per soggetto e genere); e che la provenienza dei metadati generati dall'IA e i meccanismi di governance sono raramente valutati nelle implementazioni pubblicate. L'articolo propone un'architettura a cinque livelli – acquisizione e cattura, descrizione e valutazione potenziate dall'IA, affidabilità e provenienza, individuazione e accesso, e governance – che colloca l'elaborazione del linguaggio naturale, la visione artificiale, la classificazione basata sull'apprendimento automatico e i modelli linguistici di grandi dimensioni all'interno di un quadro*

"human-in-the-loop" (con l'intervento umano) conforme ai principi archivistici di autenticità, affidabilità, integrità e usabilità, e introduce livelli di automazione per singolo campo come principio di progettazione. Le evidenze sperimentali e delle interviste indicano che la qualità del riconoscimento e della descrizione dipende dall'adattamento alla scrittura e alla collezione e che l'esperienza diretta con output imperfetti dell'IA riduce la fiducia e la disponibilità d'uso dei partecipanti. Viene inoltre definito un protocollo di validazione empirica mediante studi di caso multipli.

Conclusioni/risultati: *L'intelligenza artificiale può aumentare in modo sostanziale l'efficienza e l'accessibilità della descrizione, della valutazione e del recupero dei documenti d'archivio, ma la sua integrazione responsabile richiede strutture di governance esplicite, meccanismi di trasparenza, il trattamento dei materiali connessi all'IA come documenti a pieno titolo e una ridefinizione del ruolo professionale dell'archivista come coordinatore e controllore dei processi automatizzati, piuttosto che come semplice esecutore delle funzioni archivistiche. Poiché il modello è concettuale e basato sulle evidenze ma non validato empiricamente, l'articolo indica negli studi di caso multi-istituzionali la priorità per la ricerca futura.*

Parole chiave: *intelligenza artificiale, archivistica, gestione dei documenti, archivi digitali, descrizione archivistica, human-in-the-loop, IA affidabile.*

1. INTRODUCTION

Archival institutions worldwide are confronted with an unprecedented growth in the volume, variety and velocity of the records they are expected to appraise, describe, preserve and make accessible. Government agencies, universities, corporations and civil-society organisations now generate born-digital material – e-mail, databases, social media content, geospatial data and multimedia – at a scale that traditional, labour-intensive archival methods struggle to absorb. This growth has produced what is often called the "digital backlog" or "dark archives" problem, in which vast quantities of potentially valuable records remain unprocessed, undescribed and therefore effectively inaccessible to researchers, citizens and the institutions themselves.

Artificial intelligence (AI), understood broadly as the family of computational techniques that includes machine learning (ML), natural language processing (NLP), computer vision (CV) and, more recently, large language models (LLMs), has emerged as one of the principal responses to this challenge. Over the past decade, archival scholars and practitioners have experimented with AI-assisted transcription and handwriting recognition (Muehlberger et al., 2019), automated subject indexing, sensitivity review (Baron, Sayed and Oard, 2022; Jaillant and Caputo, 2022), entity extraction and predictive appraisal, generally reporting encouraging results in terms of processing speed and coverage. A recent systematic review identified 45 eligible studies published between 2011 and 2023 (Shinde et al., 2025); another, focused on a single handwriting-recognition platform, collected 381 papers published between 2015 and 2020 (Nockels et al., 2022); and the arrival of generative LLMs has since opened a new and fast-moving strand of experimentation in description and cataloguing (Abele et al., 2025; Nair et al., 2025; Saccucci and Potter, 2025). The result is a rapidly maturing but still fragmented research landscape.

Despite this proliferation of individual applications, the literature has so far offered comparatively little in the way of an integrated architectural vision that shows how these discrete AI tools could be combined into a coherent, standards-based archival system, and how such a system relates to established archival theory and professional ethics. Existing overviews organise the field through the records continuum (Colavizza et al., 2021) or by technique and records-manage-

ment function (Shinde et al., 2025), and the former explicitly calls for a more structural, yet critically aware, integration of AI into archival systems and practice. Practitioner studies, in turn, show that the main obstacles are as much organisational, ethical and skills-related as technical (Jaillant, 2022; Cushing and Osti, 2023). This article addresses the resulting gap. It asks three interrelated questions: (RQ1) what does the current literature reveal about the maturity, the evidence base and the risks of AI across the main archival functions; (RQ2) what would a coherent, end-to-end "intelligent archival ecosystem" look like if designed from first principles of archival science rather than purely from a technological perspective, and what governance and human-oversight mechanisms are necessary to keep it trustworthy, transparent and consistent with the professional and ethical obligations of archivists; and (RQ3) to what extent do publicly reported implementations and policy responses document the safeguards that such an architecture requires?

The article makes four contributions. First, it provides an updated, structured review of the field that integrates scholarly work, practitioner studies, standards and recent official guidance from national archives, together with a cross-study analytical synthesis of eleven empirical evaluations. Second, it proposes a five-layer reference architecture derived from that review. Third, it tests the architecture against six publicly reported implementations and policy responses through a documentary case analysis, which yields a concrete design principle – field-level automation tiers – and identifies where the published evidence is thinnest. Fourth, it specifies a protocol through which the architecture can be validated empirically in future multiple-case research.

The remainder of the article is structured as follows. Section 2 reviews the theoretical and standards foundations, synthesises the literature by archival function and analyses the empirical evidence across studies. Section 3 explains the research design. Section 4 presents the core architectural contribution: a five-layer model of the intelligent archival ecosystem. Section 5 analyses six documented cases against the architecture. Section 6 discusses the opportunities and risks associated with the model, considers how the professional role of the archivist is likely to evolve, and sets out the limitations of the study and a protocol for empirical validation. Section 7 concludes.

2. LITERATURE REVIEW

2.1 REVIEW APPROACH

The review follows a structured narrative design. Searches were carried out in October 2026 using combinations of the terms “artificial intelligence”, “machine learning”, “large language model”, “archives”, “records management”, “archival description”, “appraisal”, “sensitivity review” and “digital preservation” across publisher platforms, institutional repositories, preprint servers and the websites of national archives, library bodies and standards organisations. Reference lists of earlier reviews were used for citation chasing. Sources were retained if they (i) addressed AI techniques, their governance or their professional implications in archival, records-management or closely analogous library and web-archiving contexts; (ii) were peer-reviewed, issued by an institution or standards body, or presented at a recognised professional congress; and (iii) were published mainly between 2019 and 2026, with earlier works retained where they are foundational, as with the records continuum, OAIS and the ICA standards. Commercial and blog content was used only as practitioner commentary, and preprints were retained only where they report original evaluations and are flagged as such in the reference list. The review is not a PRISMA-type systematic review: sources were not formally quality-appraised, only English-language material was considered, and the sample inevitably over-represents institutions in North America, Europe, Australia and East Asia. These limitations are discussed in Section 6.4.

2.2 THEORETICAL AND STANDARDS FOUNDATIONS

Contemporary archival theory has moved away from the linear “life-cycle” model, in which records pass sequentially through creation, active use, semi-active storage and final disposition or permanent preservation, toward continuum-based thinking, in which the archival, evidential and contextual qualities of a record are understood as being constructed and reinforced continuously, from the moment of creation onward, rather than assigned only at the point of transfer to an archive. This shift matters for the design of AI-powered archives because it implies that automated tools should not be conceived merely as back-end processing aids applied to inert material long after creation; ideally, they should be capable of operating at multiple points along the continuum, from records capture in active business systems through to long-term preservation and reuse.

A second theoretical anchor is the Open Archival Information System (OAIS) reference model, which decomposes archival functions into ingest, archival storage, data management, administration, preservation planning and access. First issued in 2002 and revised in 2012, OAIS now exists in a third edition, published jointly as CCSDS 650.0-M-3 (December 2024) and ISO 14721:2025 (ISO, 2025). It has become a de facto standard for digital repositories and offers a useful, technology-neutral vocabulary for describing where AI components can be inserted without disturbing the underlying responsibilities of a trusted digital repository.

Descriptive standards are also changing. The International Council on Archives' Records in Contexts (RiC) standard, whose conceptual model (RiC-CM) and ontology (RiC-O) reached version 1.0 in late 2023 and whose application guidelines were completed in 2025, reconciles and replaces ISAD(G) together with three further ICA standards (ICA-EGAD, 2023). RiC describes records, agents and activities as interrelated entities that can be published as linked data through an OWL ontology, rather than as a single hierarchy, and Dancy (2024) cautions that its adoption depends on software that largely did not exist when it was released. For AI-assisted description this matters in two ways: automatically extracted entities and relations need a target structure richer than the fonds-to-item hierarchy, and any AI-generated description must remain interoperable with the standards that archival institutions are now adopting. ISO 15489-1 continues to supply the records-management principles against which both are judged.

The InterPARES Trust AI project (also referred to as "I Trust AI"), the fifth phase of the long-running InterPARES research programme, is a useful contemporary reference point. Its explicit goal is to design, develop and leverage artificial intelligence to support the continuing availability and accessibility of trustworthy public records across an international, multi-sector partnership. According to its preliminary results, the partnership aims to identify and develop AI technologies for critical records and archives challenges, to determine the benefits and risks of employing them, to ensure that archival concepts and principles inform the development of responsible AI, and to validate outcomes through case studies and demonstrations (Frontoni et al., 2022). The project's framing – trustworthiness as the organising value against which AI applications in archives must be measured – is adopted as a guiding principle for the architecture proposed in this article.

2.3 PRIOR REVIEWS OF AI IN ARCHIVES

Colavizza et al. (2021) offer one of the most cited overviews of the field. Organised through the lens of the records continuum model, their survey identifies four broad themes: theoretical and professional considerations, the automation of recordkeeping processes, the organisation of and access to archives, and novel forms of digital archives. Among the emerging directions they highlight are the application of recordkeeping principles to the data and processes that power AI itself, and a more structural, yet critically aware, integration of AI into archival systems and practice. Shinde et al. (2025) provide the most recent systematic review located for this study. Analysing 45 papers published between 2011 and 2023 against predetermined criteria, they map key AI techniques onto records-management functions, with particular attention to retention, disposition and appraisal, and they emphasise the need for closer collaboration between AI experts and archival professionals.

Two further contributions complement these overviews. Nockels et al. (2022) present the first overarching review of HTR in published research, coded from 381 papers dated 2015–2020 that mention the Transkribus platform; they find that it features primarily in archival and library science publications, with a long tail of disciplines from history and computer science to law and education, and they describe HTR as a mature machine-learning tool that is becoming integrated into digitisation workflows even though research on how it changes the information environment is scant. At a more conceptual level, the computational archival science (CAS) agenda argues that archival theories and methods developed for paper records have become disconnected from the computational methods through which digital records are now produced and consumed, and that archival practice and computational science can each benefit from the other (Marciano et al., 2018). Rolan et al. (2019), whose definition of AI later practitioner studies adopt (Cushing and Osti, 2023), had earlier examined the place of AI in the archive.

These works differ in scope – the records continuum (Colavizza et al., 2021), records-management functions (Shinde et al., 2025), a single tool (Nockels et al., 2022) and a disciplinary agenda (Marciano et al., 2018) – and none was designed to yield a deployable architecture. They also pre-date much of the most recent work: Colavizza et al. (2021) precede the generative-AI wave, and the corpus of

Shinde et al. (2025) ends in 2023. The present review therefore concentrates on sources from 2019 to 2026 and, in the sections that follow, organises them by archival function and then analyses the empirical studies across functions, so that the evidence can feed directly into the design of the architecture.

2.4 EVIDENCE BY ARCHIVAL FUNCTION

Capture and transcription. Computer-vision techniques support optical character recognition of handwritten and printed material, image classification of photographic collections, and the automated extraction of structured information from forms and registers. The most thoroughly documented application is handwritten text recognition (HTR). Muehlberger et al. (2019) report on the Transkribus platform, developed in the EU-funded READ project, and conclude that HTR has become a usable technology that, combined with mass digitisation, can generate accurate transcripts of archival manuscripts; they also identify funding and resourcing as outstanding issues. Nockels et al. (2022) document the platform's diffusion through archival and library research, and Spina (2023) examines how HTR and a general-purpose language model can be combined in the transcription, correction and normalisation of manuscript correspondence. Evidence from Arabic-script collections shows both the promise and the data dependency of the approach. Lucas, Salah and Vidal-Gorène (2022) note that the volumes of training data available for Latin scripts are not yet achievable for Arabic scripts, and show that a word-based neural approach can reach an error rate below 5% with only ten manually transcribed pages of Maghribī manuscripts; yet a base model applied without adaptation to six out-of-domain manuscripts produced character error rates of roughly 12% to 30%, which fell to an average of about 7.7% after fine-tuning on around ten pages. Chan et al. (2024) report a character error rate of 8.6% for a transformer-based model on the largest public dataset of historical handwritten Arabic, a 51% improvement over the best baseline in the literature. HTR thus turns digitised images into searchable text, which in turn makes later description and access tasks feasible, but its accuracy depends on adaptation to the script, the hand and the collection.

Description and metadata generation. NLP techniques – including named-entity recognition, topic modelling and, increasingly, transformer-based language models – are used for automatic subject indexing, summarisation and sensitivity classification of textual records. Generative LLMs have been explored for

drafting catalogue descriptions, generating finding-aid summaries and answering natural-language reference questions, and the first controlled evaluations are now available. Nair et al. (2025) used GPT-4o to generate titles and abstracts for 112 web-archive (WARC) files in Web Archive Singapore and report a 99.9% reduction in metadata-generation costs, but also that human-curated metadata remained superior in quality and that content inaccuracies, hallucinations and translation problems were persistent; they conclude that LLMs should complement rather than replace human cataloguers. In the Library of Congress's Exploring Computational Description (ECD) experiment, several models, including GPTs and open-source LLMs, were tested on e-book data and human-in-the-loop workflows were prototyped. Saccucci and Potter (2025) report accuracy of up to 99% for structured fields such as title and author but below 50% for complex fields such as subject and genre, and conclude that performance did not yet meet the 95% quality threshold required for full automation. Other studies point in complementary directions. Groppe et al. (2025) propose an agentic system that combines several LLMs and report that this federated approach outperformed single-model solutions in metadata quality and reliability on a real-world archival dataset. Pignataro and de Souza Neto (2025), working in Portuguese-language archival practice, tested ChatGPT on the description of digital records according to ISAD(G), reporting an average accuracy of 92.04% and suggesting that human supervision could raise quality further. For photographic material, Abele et al. (2025) asked a vision-language model (InternVL2) to catalogue 40 archaeological photograph mounts and had 139 participants, 29 of them archive and archaeology experts, compare the output with expert descriptions: participants could tell the two apart at a moderate level, expert descriptions were rated more accurate and more useful, and OCR errors and hallucinations limited the perceived quality of the machine output. Parallel work in library cataloguing has examined how conversational AI systems might be used to assign subject headings, again with mixed results regarding accuracy and the continuing need for professional review (Brzustowicz, 2023). Deep-learning approaches applied specifically to archive management report measurable gains in processing throughput and retrieval accuracy compared with manual or rule-based methods (Li and Wang, 2025), and similar intelligent-technology applications are reported in document and records-management contexts more broadly (He, 2024).

Appraisal, selection and retention. Machine-learning classification, in both supervised and unsupervised modes, is applied to large-scale appraisal and retention-scheduling tasks, where archivists label a representative sample of records so that an algorithm can propose values, sensitivity levels or disposition categories for the remainder of a much larger corpus; unsupervised clustering and topic modelling are used to surface latent thematic structure in undescribed collections. Shinde et al. (2025) single out retention, disposition and appraisal as areas in which rapidly expanding content creates acute pressure. At The National Archives (UK), existing AI-powered tools have been tested for digital selection and sensitivity review, and new approaches, such as topic modelling, have been developed to discover latent topics across a corpus (Jaillant, Aske and Caputo, 2025). Within the CAS tradition, Marciano et al. (2018) present motivating case studies ranging from computational linguistics to graph analytics and computational finding aids, and Hodel (2022) discusses supervised and unsupervised machine-learning approaches to textual entities. Practitioner commentary similarly presents born-digital volumes that are too large for exhaustive manual review as the central appraisal challenge (Note, n.d.).

Sensitivity review and access. Because privacy, copyright and data-protection constraints keep most born-digital collections closed to users, sensitivity review has become a prominent application. Jaillant and Caputo (2022) argue that AI can help unlock such "dark" archives, for example by identifying sensitive information so that the remainder can be released, but they warn of inherent errors, black-box algorithms, and risks of bias and of fake or partial information, and they stress fairness, accountability and transparency. Baron, Sayed and Oard (2022) created a public annotated test collection for the deliberative process privilege exemption under the US Freedom of Information Act and found that text classifiers could often reliably detect even difficult cases when trained and evaluated on annotations from the same reviewer, whereas differences in reviewer interpretation, in record custodians and in the topics of training and test records posed additional challenges.

Professional trust, ethics and skills. Practitioner studies show that the barriers to adoption are not only technical. In 21 interviews with 26 archivists, librarians and other professionals, Jaillant (2022) identifies technological obsolescence,

copyright and permissions, data protection of sensitive material, the lack of a market for born-digital records, and gaps in scale and skills as key obstacles to access, and she proposes the ethical use of AI as one possible solution. Cushing and Osti (2023), drawing on focus groups with archival and digital-preservation practitioners, identify four themes: fitting AI into day-to-day practice, the responsible use of the technology, managing expectations about adoption, and bias. Jaillant and Rees (2023) similarly find that professional trust in AI tools among civil servants, archivists and academics is closely linked to shared codes of ethics and to perceived accountability, rather than to technical performance alone. Within InterPARES Trust AI, Arias Hernández and Rockembach (2025) derive a competency and AI-literacy framework from literature reviews, interviews with archival professionals who have applied AI to the processing of records, and validation workshops; the competencies include basic AI literacy, practical AI skills, data-related skills, tool testing and evaluation, adaptation of AI to workflows, and collaboration with IT developers, and the authors also flag algorithmic bias and data privacy as significant ethical concerns (see also Arias Hernández, Fewster and Penniman, 2024). Abele et al. (2025) add experimental evidence: after judging machine- and expert-written descriptions, participants' stated willingness to use AI tools (from a mean of 5.43 to 5.09 on a seven-point scale) and their trust in AI (from 3.86 to 3.66) declined, and experts reported lower willingness to use AI than non-experts, with concerns centred on preservation responsibility rather than technical performance. These patterns echo wider findings that perceived risks and opportunities predict willingness to use AI (Schwesig et al., 2023) and that people may avoid algorithms after seeing them err (Dietvorst, Simmons and Massey, 2015). In the journal Atlanti, Milovanović (2025) frames adoption as a set of questions that determine whether AI tools are feasible, sustainable and understandable in archival practice.

Accountability and paradata. A conceptual strand addresses how the processing performed by AI can itself be documented. Bunn (2020) examines explainable AI from a recordkeeping perspective, in contexts in which transparency is important. The InterPARES Trust AI group has proposed the concept of paradata – information about the procedures and tools used to create and process information resources, together with information about the persons carrying out

those procedures – as a way to ensure that AI-driven systems honour the archival ethic from the outset and to evaluate off-the-shelf automation (Davet et al., 2022). Davet, Hamidzadeh and Franks (2023) review the theoretical background of paradata, the documentation produced in recent machine-learning experiments, the factors that condition its use, and data structures for representing and displaying paradata to different stakeholder groups.

Records created by AI and the regulatory environment. A distinct strand, largely absent from earlier reviews, concerns the status of AI-related materials as records in their own right. The US National Archives and Records Administration (NARA, 2026) holds that the use of AI platforms does not in itself create federal records, but that prompts, outputs, audit trails, procurement and operational documentation, copies of agency records used to train AI, and algorithms or models may be records when, for example, they are relied upon in decision-making, used for official business or captured in agency systems; it also notes that existing general records schedules do not yet cover AI policy documents or bodies of data kept solely for training AI. The National Archives of Australia (n.d.) advises that AI systems that create records are generally not designed as recordkeeping systems and lack adequate records-management functionality, and that records evidencing agency business involving AI, including decisions and the reasoning behind them, must be retained. NARA's own AI compliance plan describes an AI governance board, a use-case inventory and internal guidance on generative AI (NARA, 2025). At the level of regulation and risk management, the European Union's Artificial Intelligence Act (European Union, 2024) establishes a risk-based regime that includes provisions on human oversight, automatic logging and AI literacy, while the NIST AI Risk Management Framework (NIST, 2023) organises AI risk management into four functions: govern, map, measure and manage.

2.5 CROSS-STUDY ANALYTICAL SYNTHESIS

To move from description to analysis, the eleven empirical studies that report original evaluations, interviews or experiments were coded for task, metric, headline result and stance on human oversight. Table 1 summarises the coding, and six analytical observations follow.

Table 1. Cross-study synthesis of eleven empirical studies

Study	Task and material	Metric	Headline result	Stance on human oversight
Lucas et al. (2022)	HTR of Arabic Maghribī manuscripts	Character error rate (CER)	Below 5% with ten transcribed pages; base model 12–30% on out-of-domain manuscripts	Iterative transcription, correction and re-training
Chan et al. (2024)	HTR of historical handwritten Arabic	CER	8.6% on the largest public historical dataset; 51% better than best baseline	Not addressed in the abstract
Nair et al. (2025)	GPT-4o titles and abstracts for 112 WARC files	Cost; human evaluation	99.9% cost reduction; human metadata of higher quality; hallucinations	LLM as complement to cataloguers
Saccucci and Potter (2025)	LLM-generated catalogue fields for e-books	Field-level accuracy	Up to 99% for title and author; below 50% for subject and genre	Human-in-the-loop; 95% threshold for full automation
Pignataro and de Souza Neto (2025)	ChatGPT description of digital records under ISAD(G)	Accuracy	Average accuracy of 92.04%	Human supervision could raise quality
Groppe et al. (2025)	Multi-LLM metadata generation for archival materials	Metadata quality and reliability	Federated ensemble outperformed single models	Not addressed in the abstract
Abele et al. (2025)	Vision-language model descriptions of 40 archaeological photograph mounts; 139 participants	Detection sensitivity; ratings; trust	Detectable at a moderate level; expert text rated more accurate and useful; trust and willingness fell after exposure	AI drafts, subordinate to human verification
Baron, Sayed and Oard (2022)	Classification of deliberative-process material (FOIA)	Classifier performance	Reliable with same-reviewer annotations; varies by reviewer, custodian and topic	Advisory classification; reviewer variance documented
Jaillant (2022)	21 interviews with 26 professionals	Thematic analysis	Obstacles: obsolescence, copyright, data protection, scale, skills	Ethical use of AI; collaboration across disciplines
Cushing and Osti (2023)	Focus groups with archival practitioners	Thematic analysis	Four themes: day-to-day fit, responsible use, expectations, bias	Responsible use central
Arias Hernández and Rockembach (2025)	Interviews and validation workshops (InterPARES Trust AI)	Competency and literacy frameworks	Literacy, tool testing and collaboration with IT developers	Competence needed to oversee AI

Source: author's coding of the studies cited. Abele et al. (2025), Chan et al. (2024), Groppe et al. (2025) and Lucas et al. (2022) are preprints.

(1) Performance follows the interpretive demand of the task rather than the technology. The highest figures are reported where the task is recognition or extraction under controlled conditions: up to 99% accuracy for title and author fields (Saccucci and Potter, 2025), character error rates below 5% after adaptation on ten pages (Lucas et al., 2022) and 8.6% on a large historical Arabic benchmark (Chan et al., 2024). Interpretive tasks fall much lower: subject and genre assignment stayed below 50% in the same Library of Congress experiment, and

sensitivity classification performed well only when the annotator was consistent (Baron, Sayed and Oard, 2022). The studies therefore suggest a gradient from recognition, through extraction and interpretation, to judgement, in which each step demands more human involvement.

(2) Adaptation to the collection often matters as much as the choice of model. Fine-tuning on about five pages halved the average character error rate of an Arabic HTR model (from 21.24% to 9.99%) in Lucas et al. (2022); customisation for historical Arabic produced a 51% improvement over the best baseline in Chan et al. (2024); combining several LLMs outperformed single models in Groppe et al. (2025); and Abele et al. (2025) identify domain-specific fine-tuning as a route to fewer hallucinations. Where adaptation depends on verified examples, the archivist's verified samples become the scarce input on which automation quality depends.

(3) Evaluation metrics are heterogeneous and can mislead. The studies report character error rates, accuracy percentages, cost reductions, F1 scores and human detection sensitivity, so results cannot be compared directly. Lucas et al. (2022) also show that character error rate can penalise correct output when transcription conventions differ from the ground truth and that a mean can conceal the distribution of errors across a page, since half of the lines of one manuscript had error rates below 10% although its mean was 12.24%. Without harmonised, ground-truth-based protocols, institutions cannot tell whether a published result will transfer to their own records.

(4) Human experience diverges from technical metrics. In the only controlled experiment among the eleven studies, in which 29 of 139 participants were archive or archaeology experts, machine descriptions rated higher for accuracy were harder to identify as machine-made, yet hands-on exposure lowered willingness to use AI and trust in it, and experts were less willing than non-experts (Abele et al., 2025). This fits the practitioner findings that responsible use, managing expectations and bias dominate adoption debates (Cushing and Osti, 2023) and that competence in testing and adapting tools is a core skill (Arias Hernández and Rockembach, 2025). It should be read with caution, because the sample of Abele et al. was predominantly young non-experts (110 of 139) and the study is a preprint. Trust is therefore better treated as a managed design variable than as a by-product of accuracy.

(5) Studies converge on human-in-the-loop designs but differ on what the human does. Draft-and-review (Nair et al., 2025; Abele et al., 2025), suggestion with confidence scores (Saccucci and Potter, 2025), human supervision of ChatGPT output (Pignataro and de Souza Neto, 2025) and calibrated review of sensitivity flags (Baron, Sayed and Oard, 2022) are all variants of the same principle, but only the Library of Congress states an explicit numerical threshold for automation (95%). This gap motivates the field-level automation tiers developed in Section 5.3.

(6) Accountability documentation lags behind the conceptual literature. The paradata literature argues that information about the procedures, tools and persons involved in AI processing should be captured with the records (Davet et al., 2022; Davet, Hamidzadeh and Franks, 2023). In the empirical studies, processing details appear mainly in research methods or supplements – Abele et al. (2025), for example, publish prompts and temperature settings as supplementary material – rather than as records attached to the descriptions produced. How far institutions capture such evidence in routine practice is not shown by the studies, which is a further argument for the empirical protocol in Section 6.4.

2.6 SYNTHESIS AND RESEARCH GAPS

Table 2 summarises the reviews, conceptual frameworks and policy sources that frame the empirical evidence.

Table 2. Reviews, conceptual frameworks and policy sources informing the architecture

Source	Type and scope	Key contribution	Relevance to architecture
Colavizza et al. (2021)	Survey of the archives-and-AI literature, organised by the records continuum	Four themes; calls for applying recordkeeping principles to AI data and processes and for structural, critically aware integration of AI	Theoretical frame; Layers 3 and 5
Shinde et al. (2025)	Systematic review of 45 papers (2011–2023)	Maps AI techniques to records-management functions; stresses collaboration between AI experts and archivists	Layer 2; Layer 5
Nockels et al. (2022)	Systematic review of 381 papers (2015–2020) on one HTR platform	HTR is a mature tool, concentrated in archival and library science, spreading across disciplines	Layer 1
Marciano et al. (2018)	Conceptual chapter with motivating case studies (CAS)	Argues for closer links between archival theory and computational methods	Theoretical frame; Layers 2 and 5

Source	Type and scope	Key contribution	Relevance to architecture
Bunn (2020); Davet et al. (2022, 2023)	Conceptual and review papers on explainability and paradata	AI processing should be documented as paradata to support accountability	Layer 3
NARA (2025, 2026); National Archives of Australia (n.d.)	Official guidance and governance plan	AI materials may be records; scheduling and capture requirements; governance structures	Layers 3 and 5
European Union (2024); NIST (2023)	Regulation and risk-management framework	Human oversight, logging and AI literacy; govern, map, measure and manage	Layer 5

Source: author's synthesis of the literature reviewed in Sections 2.2–2.4.

Four gaps emerge from the review and the cross-study synthesis. First, no integrated reference architecture links AI techniques, archival standards, human oversight and governance (G1). Second, much of the evidence on operational performance comes from single-institution experiments, and results vary sharply by task: structured fields can be extracted with high accuracy, whereas interpretive tasks such as subject and genre assignment, sensitivity judgement and appraisal remain far from automation-grade, and studies use incompatible metrics (G2). Third, the provenance of AI-generated metadata, the paradata proposed in the conceptual literature and the management of the AI system's own records are discussed mainly in conceptual and policy documents rather than evaluated in implementations (G3). Fourth, evidence on non-Latin scripts exists – notably for Arabic handwriting (Lucas et al., 2022; Chan et al., 2024) – but it concerns the recognition of manuscripts rather than the description, appraisal and access workflows of records-management institutions, and the sources consulted say little about institutions with limited technical resources (G4). The architecture proposed in Section 4 responds directly to G1–G3; G2 and G4 are carried forward into the validation protocol in Section 6.4.

3. METHODOLOGY

This article adopts a conceptual, design-oriented approach, supported by a structured literature review and a documentary case analysis, rather than primary empirical data collection. It follows the logic of design-science reasoning, in which existing theory and a synthesis of prior literature are used to propose an artefact – in this case, a layered architectural model – intended to solve a practical problem identified in the field, namely the absence of an integrated, standards-consistent framework for combining discrete AI tools into a single archival system.

The research design has four components. The first is the structured narrative review described in Section 2.1, which provides the evidence base. The second is a cross-study analytical synthesis (Section 2.5), in which eleven empirical studies were coded for task, metric, headline result and stance on human oversight, and which identifies the research gaps. The third is the design of the architecture itself (Section 4), guided by four criteria: (1) fidelity to established archival principles, so that automated components reinforce rather than undermine the authenticity, reliability, integrity and usability of records; (2) modularity, so that individual AI components (for example an OCR engine or a classification model) can be adopted, replaced or upgraded independently without redesigning the whole system; (3) explicit accountability, so that every automated decision affecting appraisal, description or access can be traced to a human-defined policy and, where necessary, reviewed or reversed by a qualified archivist; and (4) evidential grounding, so that each layer can be traced to documented practice or to a documented risk in the literature.

The fourth component is a documentary case analysis (Section 5). Cases were selected purposively against four criteria: the implementation, experiment or policy response is publicly documented; it concerns archives, records management or closely analogous library and web-archiving contexts; it was reported between 2019 and 2026; and, taken together, the cases cover different layers of the architecture and different types of institution. Six cases met these criteria and were retained. Each was coded against five dimensions derived from the architecture: the layer or layers addressed; the AI technique and scope; the reported evidence of performance or effect; the reported limits; and the safeguards or oversight arrangements described. The data are secondary – published articles, institutional reports, presentations and official guidance – and no primary data were collected.

The resulting model should be read as a conceptual reference architecture rather than a technical specification; its purpose is to give archival institutions, standards bodies and software developers a shared vocabulary and structure for planning AI adoption, and to provide researchers with a framework that can subsequently be tested, refined or challenged through empirical case studies (Section 6.4).

This approach carries evident limitations. The review is structured but not exhaustive, and both the synthesis and the case coding rest on published reports,

which may omit undocumented practice; several of the empirical studies are preprints that have not completed peer review. Because the study is not based on primary empirical data collection such as interviews, surveys or system evaluations, the model cannot claim to be validated against real institutional practice; its value lies instead in synthesising a dispersed literature into a single, theoretically grounded and evidence-informed proposal that can orient future empirical work. These limitations are revisited in Section 6.4.

4. TOWARD AN INTELLIGENT ARCHIVAL ECOSYSTEM: A LAYERED ARCHITECTURE

Building on the theoretical foundations and the evidence reviewed above, this section proposes a five-layer architecture for an intelligent archival ecosystem. The layers correspond broadly to the archival functions identified in OAIS and continuum-based theory, but each is reinterpreted to specify where and how AI components can be embedded, and what safeguards are required at each point. Table 3 summarises the architecture; the subsections that follow explain each layer.

Table 3. The five-layer architecture of the intelligent archival ecosystem

Source	Type and scope	Key contribution	Relevance to architecture
Colavizza et al. (2021)	Survey of the archives-and-AI literature, organised by the records continuum	Four themes; calls for applying recordkeeping principles to AI data and processes and for structural, critically aware integration of AI	Theoretical frame; Layers 3 and 5
Shinde et al. (2025)	Systematic review of 45 papers (2011–2023)	Maps AI techniques to records-management functions; stresses collaboration between AI experts and archivists	Layer 2; Layer 5
Nockels et al. (2022)	Systematic review of 381 papers (2015–2020) on one HTR platform	HTR is a mature tool, concentrated in archival and library science, spreading across disciplines	Layer 1
Marciano et al. (2018)	Conceptual chapter with motivating case studies (CAS)	Argues for closer links between archival theory and computational methods	Theoretical frame; Layers 2 and 5
Bunn (2020); Davet et al. (2022, 2023)	Conceptual and review papers on explainability and paradata	AI processing should be documented as paradata to support accountability	Layer 3
NARA (2025, 2026); National Archives of Australia (n.d.)	Official guidance and governance plan	AI materials may be records; scheduling and capture requirements; governance structures	Layers 3 and 5
European Union (2024); NIST (2023)	Regulation and risk-management framework	Human oversight, logging and AI literacy; govern, map, measure and manage	Layer 5

Source: author's design, derived from the review in Section 2 and the case analysis in Section 5.

4.1 LAYER 1 – INGESTION AND CAPTURE

The first layer covers the point at which records enter the ecosystem, whether through direct transfer from an active records-management system, bulk digitisation of physical holdings, or harvesting of web and social-media content. AI contributes here through automated file-format identification, duplicate and near-duplicate detection, optical character recognition and handwriting recognition, and preliminary quality-control checks (for example, flagging corrupted files or illegible scans for human review). Capturing rich technical and administrative metadata automatically at this stage – rather than reconstructing it later – is essential, because continuum theory holds that the evidential value of a record depends on context established as close as possible to the moment of creation or transfer. Because the quality of every downstream process depends on the quality of capture, this layer also stores the recognition confidence values produced by OCR and HTR engines alongside the text. Evidence from handwritten text recognition shows that this layer can already operate at institutional scale, although funding and resourcing remain recurring constraints (Muehlberger et al., 2019), and the Arabic-script studies indicate that recognition models usually need adaptation to each collection's script and hands, which makes a small, archivist-verified sample a standing input to this layer (Lucas, Salah and Vidal-Gorène, 2022).

4.2 LAYER 2 – AI-AUGMENTED DESCRIPTION, CLASSIFICATION AND APPRAISAL

The second layer is where the greatest number of current AI applications are concentrated. NLP-based entity recognition and topic modelling can generate draft subject terms, proposed access points and preliminary summaries consistent with descriptive standards such as ISAD(G) and, increasingly, Records in Contexts, whose entity-and-relation structure offers a natural target for automatically extracted agents, activities and relationships; computer-vision classifiers can group visual material by content type, estimated date or format; and supervised machine-learning models, trained on samples labelled by archivists, can propose retention or disposition recommendations for large, undifferentiated bodies of records, addressing what Note (n.d.) and others describe as the central appraisal challenge posed by born-digital volumes that are simply too large for exhaustive manual review. Crucially, in the proposed architecture every AI-generated de-

scription, classification or appraisal recommendation is marked as provisional metadata until reviewed, and confidence scores are exposed to the archivist so that low-confidence outputs are prioritised for manual verification rather than silently accepted.

The evidence reviewed in Section 2.4 shows that automation quality varies sharply from one metadata field to another. Layer 2 is therefore not designed as a single switch between manual and automatic description; Section 5.3 derives a system of field-level automation tiers that determines, for each field and decision type, how much human review is required. The cross-study synthesis (Section 2.5) also shows that adaptation to the collection can matter as much as model choice, so Layer 2 includes an adaptation loop in which archivist-verified descriptions are fed back as training or prompt examples; where several models are available, their outputs can be compared or combined, as in the ensemble approach of Groppe et al. (2025).

4.3 LAYER 3 – TRUST, PROVENANCE AND AUTHENTICITY

The third layer addresses a concern that is specific to archival science and largely absent from generic AI-system design: the need to preserve, and be able to demonstrate, the authenticity and integrity of records over time, including any transformation introduced by AI processing itself. This layer maintains a provenance log recording which AI model, version and configuration produced or modified a given piece of metadata, when, and under whose authorisation, in line with the trustworthiness principles emphasised by the InterPARES Trust AI project. Digital-provenance and content-authenticity standards emerging from initiatives such as the Coalition for Content Provenance and Authenticity offer one possible technical basis for this layer, alongside cryptographic fixity checks already common in digital-preservation systems. The objective is that an AI-assisted description can always be traced back to its origin and distinguished from human-authored description, preserving the chain of custody that underpins archival authenticity. In the vocabulary of the InterPARES Trust AI group, this log is a form of paradata: information about the procedures and tools used to process records and about the persons responsible (Davet et al., 2022; Davet, Hamidzadeh and Franks, 2023).

Recent official guidance adds a second requirement. If prompts, outputs, audit trails, model documentation and copies of records used to train or tune models may themselves be records (NARA, 2026; National Archives of Australia, n.d.),

then the processing evidence generated by the ecosystem must be managed under records schedules rather than left to the retention settings of a vendor platform. NARA observes that information created by AI and passively retained by a third-party platform is not necessarily "received" by an agency until it is captured in an agency system, which argues for capturing processing evidence inside the archive's own environment.

4.4 LAYER 4 – DISCOVERY AND ACCESS

The fourth layer is user-facing and covers search, retrieval and reference services. AI contributes through semantic and natural-language search that goes beyond exact keyword matching, automated sensitivity classification that helps identify material requiring redaction or restricted access before release, and conversational reference assistants capable of guiding researchers through finding aids. Because access decisions can carry legal and ethical consequences – particularly around personal data, national security or culturally sensitive material – this layer is designed so that AI recommendations for release or restriction are advisory rather than determinative, with final release decisions remaining a documented human responsibility, consistent with the emphasis on impartial and accountable access found in the International Council on Archives' Code of Ethics. Because classifier performance has been shown to vary with the reviewer who produced the training annotations, the custodian of the records and the topic (Baron, Sayed and Oard, 2022), the layer also records each reviewer decision and calibrates models separately for different record series and custodians.

4.5 LAYER 5 – GOVERNANCE AND HUMAN OVERSIGHT

The fifth layer is not a processing stage but a cross-cutting function that governs the other four. It comprises policy documentation of which AI tools are approved for which tasks; role-based human review workflows, so that AI outputs of consequence always pass through a qualified archivist before becoming authoritative; auditing and explainability mechanisms that make it possible to inspect why a model produced a given recommendation; and periodic bias and performance evaluation of deployed models against representative samples of the collection. This layer operationalises, at the level of system design, the ethical concerns raised repeatedly in the literature regarding algorithmic bias, opacity and the risk of reproducing historical prejudices present in training data or in the archive itself.

The layer can be organised around the four functions of the NIST AI Risk Management Framework (NIST, 2023): govern (policies, roles and the register of approved tools), map (documenting each AI use case and its context before deployment, as in the use-case inventory and assessment worksheets used by NARA and the Library of Congress), measure (field-level performance, bias and drift testing) and manage (escalation, rollback and retirement of models). Where an institution operates under the European Union's Artificial Intelligence Act (European Union, 2024), the requirements for human oversight, logging and staff AI literacy map directly onto the review workflows of Layers 2 and 4, the provenance log of Layer 3 and the competency programme described in Section 6.3. Because trust falls after hands-on exposure to imperfect output and is lower among experts (Abele et al., 2025), the layer also treats trust calibration as a managed variable, through transparent presentation of confidence and provenance, supervised practice with the tools, and the AI-literacy competencies identified by Arias Hernández and Rockembach (2025).

5. DOCUMENTARY CASE ANALYSIS

5.1 CASES AND CODING

To test the plausibility of the architecture against published evidence, six cases were analysed using the procedure described in Section 3. Table 4 summarises the coding. Two points of caution apply. The Library of Congress experiment concerns library cataloguing rather than archival description, but it is the most fully reported evaluation of LLM-based description and is closely analogous to Layer 2. The sixth case combines two sets of official guidance and a governance plan because they address the same layers and are complementary.

Table 4. Coding of six documented cases against the architecture

Layer	Main AI techniques	Human role	Key safeguards	Evidence base
1. Ingestion and capture	Format identification; duplicate detection; OCR and HTR; quality-control flagging	Sets intake policy; reviews flagged items	Automatic capture of technical metadata and recognition confidence values; fixity	Muehlberger et al. (2019); Nockels et al. (2022); Lucas et al. (2022); continuum theory
2. AI-augmented description, classification and appraisal	Entity recognition; topic modelling; supervised classification; LLM summarisation; image classification	Reviews provisional metadata; labels training samples; decides on appraisal	Provisional status; confidence scores; field-level automation tiers	Nair et al. (2025); Saccucci and Potter (2025); Groppe et al. (2025); Abele et al. (2025)

Layer	Main AI techniques	Human role	Key safeguards	Evidence base
3. Trust, provenance and authenticity	Provenance logging; content-credential standards; fixity checks	Authorises models and configurations; audits logs	Processing-provenance log; AI-related materials managed as records; AI and human description distinguished	InterPARES Trust AI; Davet et al. (2023); Bunn (2020); NARA (2026); National Archives of Australia (n.d.); C2PA (2024)
4. Discovery and access	Semantic search; sensitivity classification; conversational reference assistants	Makes final release and restriction decisions	Advisory-only recommendations; documented decisions; calibration by reviewer and custodian	Baron et al. (2022); Jaillant and Caputo (2022); ICA Code of Ethics
5. Governance and human oversight (cross-cutting)	Model evaluation; bias testing; explainability tools	Sets policy; audits; maintains competencies	Register of approved tools; role-based review; periodic bias and performance evaluation; use-case inventory; AI literacy	NIST (2023); European Union (2024); Arias Hernández and Rockembach (2025); Cushing and Osti (2023)

Source: author's coding of the sources cited in the table. The coding reflects what each source reports and does not exclude undocumented practice.

5.2 CROSS-CASE FINDINGS

Four patterns emerge from the coding. First, performance depends on the interpretive demand of the task. In the two description experiments, structured or surface-level outputs (title, author, a draft abstract) were produced with high accuracy or large cost savings, whereas interpretive outputs (subject, genre) were far below the level required for automation (Nair et al., 2025; Saccucci and Potter, 2025). The same logic appears in sensitivity review, where classifiers performed well on difficult cases only when annotation was consistent (Baron, Sayed and Oard, 2022).

Second, each of the three evaluation studies of description and review positions AI as an assistant to human cataloguers or reviewers rather than a replacement: Nair et al. (2025) explicitly recommend complementarity, the Library of Congress prototypes human-in-the-loop workflows and measures output against a 95% threshold, and Baron, Sayed and Oard (2022) document reviewer variance. This convergence supports treating human review as a design requirement (criterion 3 in Section 3) rather than as an optional safeguard. The additional studies synthesised in Section 2.5 point the same way, from the draft-and-review design of the vision-language experiment of Abele et al. (2025) to the human supervision recommended by Pignataro and de Souza Neto (2025).

Third, the published evidence is concentrated in Layers 1, 2 and 4. In the sources consulted, none of the implementation reports presents an evaluation of provenance logging for AI-generated metadata (Layer 3) or of governance mechanisms

such as bias audits and model retirement (Layer 5). These layers appear mainly in policy documents: NARA and the National Archives of Australia address how AI materials become records, and NARA describes governance structures, but neither source reports on how such arrangements perform in practice, and NARA's memorandum is explicitly limited to records-management requirements, leaving AI governance, ethics, privacy and security to other instruments (NARA, 2026).

Fourth, comparability is limited. The cases differ in institutional type (national archives, library, web archive), in language and script, and in maturity, and at least three of the six are research experiments rather than production deployments. The cases therefore test the plausibility of the architecture and expose its gaps; they do not validate it.

5.3 DESIGN IMPLICATION: FIELD-LEVEL AUTOMATION TIERS

The cross-case findings suggest one refinement to the architecture. Instead of deciding, at the level of a whole collection or tool, whether description is "automated", Layer 2 should assign each metadata field and each decision type to an automation tier according to measured accuracy against an institution-defined quality threshold. Tier A (assisted, with sampling) covers fields whose measured accuracy exceeds the threshold: outputs may be accepted as provisional metadata and verified by statistical sampling. Tier B (assisted, with full review) covers fields that fall below the threshold: outputs are presented to the archivist as suggestions with confidence scores and each must be reviewed. Tier C (human decision) covers decisions with legal, ethical or historical significance – final access restrictions, destruction decisions and the description of culturally sensitive material – where AI is limited to retrieval and summarisation support.

The Library of Congress's 95% threshold for full automation (Saccucci and Potter, 2025) offers one example of how such a threshold can be expressed, although each institution would need to set its own and review it periodically, because performance changes with models, languages and collections. On the figures reported, title and author fields could approach Tier A, while subject and genre fields clearly fall into Tier B; this classification is illustrative and must be re-derived locally. The assigned tier is recorded in the provenance log (Layer 3), and the thresholds themselves are reviewed under the governance layer (Layer 5), so that the principle connects the layers rather than operating inside one of them.

6. DISCUSSION: OPPORTUNITIES, RISKS AND THE FUTURE ROLE OF THE ARCHIVIST

6.1 OPPORTUNITIES

The principal opportunity offered by an intelligent archival ecosystem, as conceived here, is scale: AI-assisted description and appraisal make it realistic to address backlogs that would otherwise remain permanently "dark", thereby improving both the accountability of institutions that create records and public access to historical and administrative memory. The 99.9% reduction in metadata-generation cost reported by Nair et al. (2025) illustrates the order of magnitude of the potential savings, although it was achieved with a loss of quality relative to human cataloguing. Secondary benefits include more consistent application of descriptive standards across large collections, faster sensitivity review that can accelerate the release of historically significant government records, and the ability to surface thematic connections across dispersed collections that would be difficult for individual archivists to identify manually.

6.2 RISKS AND ETHICAL CONSIDERATIONS

These opportunities are accompanied by significant risks, which the proposed architecture attempts to mitigate through its trust and governance layers but cannot eliminate entirely. Algorithmic bias is the most frequently cited concern in the literature: if training data reflect historical patterns of exclusion or marginalisation, AI tools risk automating and even amplifying those patterns in appraisal and description decisions, a concern archivists themselves raise in focus-group studies (Cushing and Osti, 2023). Jaillant and Rees (2023) similarly find that professional trust in AI tools among civil servants, archivists and academics is closely linked to shared codes of ethics and to perceived accountability, rather than to technical performance alone; where communication and shared standards are weak, mistrust of AI tools tends to increase even when the underlying technology functions adequately.

Experimental evidence also points to a human risk. Abele et al. (2025) found that hands-on exposure to imperfect machine descriptions lowered willingness to use AI and trust in it, and that the machine descriptions rated most accurate were the hardest to detect, which suggests that fluent output can conceal errors from reviewers. Both over-trust and under-trust are therefore design concerns, and the

architecture addresses them through confidence display, provenance and training rather than relying on reviewer vigilance alone.

A further risk concerns explainability: many contemporary machine-learning models, particularly deep-learning and large-language-model systems, function as "black boxes" whose internal reasoning is difficult to audit, which is in tension with the archival requirement that decisions affecting access and disposition be documented and defensible (Jaillant and Caputo, 2022). Over-reliance on generative AI for descriptive text carries a specific risk of factual inaccuracy or fabricated detail, documented in the Web Archive Singapore experiment (Nair et al., 2025). Variation across reviewers and custodians shows that sensitivity classifiers can encode the judgement of whoever labelled the training data (Baron, Sayed and Oard, 2022). Finally, the use of third-party platforms creates a recordkeeping risk of its own: AI-related materials that matter as evidence may sit outside the institution's control and outside its retention schedules (NARA, 2026; National Archives of Australia, n.d.).

6.3 THE ARCHIVIST AS ORCHESTRATOR

Taken together, these opportunities and risks point toward a redefinition, rather than a diminishment, of the archivist's professional role. Competency studies conducted with archivists, records managers and digital-archives researchers across several countries find that the skills increasingly required include basic AI literacy, the ability to evaluate and test AI tools before deployment, data-related skills for preparing and auditing training samples, and the capacity to collaborate effectively with information-technology developers (Arias Hernández, Fewster and Penniman, 2024; Arias Hernández and Rockembach, 2025). In the architecture proposed here, the archivist's function shifts from being the sole executor of every descriptive and appraisal task toward orchestrating, auditing and validating a partially automated workflow – setting policy at the governance layer, curating training data, interpreting confidence scores, assigning automation tiers, and retaining final authority over decisions with legal, ethical or historical significance. The Arabic HTR evidence illustrates why this matters: a few pages of expert transcription transform model performance (Lucas, Salah and Vidal-Gorène, 2022), so the archivist's verified samples are the scarce input on which automation quality depends. This reframing is consistent with calls in the professional

literature for an explicit ethical framework accompanying AI adoption in appraisal, in which transparency, auditability and human oversight are treated as design requirements rather than optional add-ons, and with practitioners' own emphasis on responsible use and on managing expectations (Cushing and Osti, 2023).

6.4 LIMITATIONS AND A PROTOCOL FOR EMPIRICAL VALIDATION

The study has four principal limitations. First, the review is structured but not exhaustive: it covers English-language sources, was not formally quality-appraised and over-represents institutions in North America, Europe, Australia and East Asia. Second, the case coding rests on published reports and summaries and on a small number of cases, and may therefore omit undocumented practice. Third, several cases are experiments or concern neighbouring sectors such as libraries and web archives, so transfer to other archival settings is uncertain. Fourth, and most important, the architecture has not been validated empirically. The next step is therefore not further conceptual elaboration but application in real institutions.

Empirical validation should proceed through a multiple-case study of three to five archival institutions that differ in size, mandate and technical maturity, including at least one that works with non-Latin-script records. The protocol has four steps. Step 1 is a baseline maturity assessment: a structured instrument based on the five layers is completed through document review and semi-structured interviews with archivists, records managers and IT staff, establishing which layers are present, partly present or absent. Step 2 is layer-specific measurement in a pilot: for Layer 1, character or word error rates, reported as medians and quartiles as well as means, and the proportion of items flagged for human review; for Layer 2, field-level precision, recall and F1 against archivist-verified descriptions, the reviewer correction rate and the time per record; for Layer 3, the share of AI-generated metadata with complete processing provenance (model, version, configuration, date and authoriser); for Layer 4, document-level precision and recall of sensitivity flags, inter-reviewer agreement and the rate of overturned recommendations; and for Layer 5, coverage of governance functions, the frequency of bias and performance audits, and staff AI-literacy outcomes. Step 3 is an evaluation of the archivist's role through focus groups or interviews after the pilot, building on the themes identified by Cushing and Osti (2023), complemented by an experimental component modelled on the

pre-registered design of Abele et al. (2025), in which practising archivists compare AI-generated and human descriptions and report perceived quality, trust and willingness to use AI. Because studies use incompatible metrics, and because character error rate can penalise correct output when transcription conventions differ and can conceal the distribution of errors (Lucas, Salah and Vidal-Gorène, 2022), all measures are taken against archivist-verified ground truth and reported with their distributions. Step 4 is cross-case synthesis, using pattern matching against the propositions of the architecture (for example, that field-level automation tiers reduce review effort without lowering accuracy) and a panel of experts to validate and refine the model. Ethical approval, data-protection impact assessment and institutional consent would be required wherever the records contain personal data.

7. CONCLUSION

This article has argued that the growing body of discrete AI applications in archival practice – spanning transcription, description, classification, appraisal and access – can usefully be organised into a single, theory-grounded architecture rather than treated as a collection of isolated pilot projects. Drawing on a structured review of mainly 2019–2026 literature, a cross-study synthesis of eleven empirical evaluations and a documentary analysis of six cases, it has shown that performance depends heavily on the interpretive demands of each task and on adaptation to the script and the collection, that published implementations consistently position AI as an assistant to human judgement, that trust is a design variable rather than a by-product of accuracy, and that provenance and governance are the least evaluated parts of current practice. The proposed five-layer model of the intelligent archival ecosystem – ingestion and capture, AI-augmented description and appraisal, trust and provenance, discovery and access, and governance – together with its field-level automation tiers, is intended to give institutions, standards bodies and researchers a shared reference framework for planning AI adoption in a manner consistent with core archival principles of authenticity, reliability, integrity and usability, and with professional codes of ethics emphasising impartial, accountable service to users.

Because the model presented here is conceptual, developed through literature synthesis and documentary analysis rather than empirical testing, the most pressing direction for future research is validation: case studies of archival institutions

that have implemented AI tools across multiple archival functions, following the protocol set out in Section 6.4, comparative evaluation of the governance mechanisms institutions actually use, and empirical measurement of bias and error rates in deployed appraisal and description models. Continued collaboration between archival scholars, computer scientists and practitioners – of the kind pursued by multi-institutional initiatives such as InterPARES Trust AI – will be essential for translating the architectural principles outlined in this article into operational, trustworthy systems that serve both the evidentiary responsibilities of archives and the public's right of access to its documentary heritage.

REFERENCE LIST

Abele, L., Anders, G., Aydın, T., Buder, J., Fischer, H., Kimmel, D. and Huff, M. (2025). *ArchiveGPT: A human-centered evaluation of using a vision language model for image cataloguing*. arXiv preprint arXiv:2507.07551.

Arias Hernández, R. and Rockembach, M. (2025). Building trustworthy AI solutions: Integrating artificial intelligence literacy into records management and archival systems. *AI & Society*, 40(6), 4265–4282.

Arias Hernández, R., Fewster, K. and Penniman, S. (2024). Artificial Intelligence and Machine Learning Competencies for the Archival Professions. In L. M. Given (ed.), *Proceedings of the Association for Information Science and Technology*, 61(1) (pgs. 36–43). New Jersey, USA: Wiley.

Baron, J. R., Sayed, M. F. and Oard, D. W. (2022). Providing more efficient access to government records: A use case involving application of machine learning to improve FOIA review for the deliberative process privilege. *ACM Journal on Computing and Cultural Heritage*, 15(1), Article 5, 1–19.

Brzustowicz, R. (2023). From ChatGPT to CatGPT: The implications of artificial intelligence on library cataloging. *Information Technology and Libraries*, 42(3).

Bunn, J. (2020). Working in contexts for which transparency is important: A record-keeping view of explainable artificial intelligence (XAI). *Records Management Journal*, 30(2), 143–153. https://doi.org/10.1108/RMJ-08-2019-0038

Chan, A., Mijar, A., Saeed, M., Wong, C.-W. and Khater, A. (2024). *HATFormer: Historic handwritten Arabic text recognition with transformers*. arXiv preprint arXiv:2410.02179.

Coalition for Content Provenance and Authenticity [C2PA]. (2024). *Content Credentials: C2PA Technical Specification*.

Colavizza, G., Blanke, T., Jeurgens, C. and Noordegraaf, J. (2021). Archives and AI: an overview of current debates and future perspectives. *ACM Journal on Computing and Cultural Heritage, 15*(1), 1–15.

Cushing, A. L. and Osti, G. (2023). "So how do we balance all of these needs?": How the concept of AI technology impacts digital archival expertise. *Journal of Documentation*, 79(7), 12–29.

Dancy, R. (2024). Waiting for RiC: The release of Records in Contexts, version 1.0. *Archivaria*, 98.

Davet, J. E., Hamidzadeh, B., Franks, P. C. and Bunn, J. (2022). Tracking the functions of AI as paradata and pursuing archival accountability. In *Archiving 2022: Expanding Connections Across Digital Cultural Heritage – Final Program and Proceedings* (pp. 83–88). Springfield, VA: Society for Imaging Science and Technology. https://doi.org/10.2352/issn.2168-3204.2022.19.1.17

Davet, J., Hamidzadeh, B. and Franks, P. (2023). Archivist in the machine: Paradata for AI-based automation in the archives. *Archival Science*, 23(2), 275–295. https://doi.org/10.1007/s10502-023-09408-8

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689, 12 July 2024.

Frontoni, E., Paolanti, M., Lauriault, T. P., Stiber, M., Duranti, L. and Abdul-Mageed, M. (2022). *Trusted data forever: Is AI the answer?* arXiv preprint arXiv:2203.03712.

Groppe, J., Marquet, A., Walz, A. and Groppe, S. (2025). *Automated archival descriptions with federated intelligence of LLMs*. arXiv preprint arXiv:2504.05711.

He, Y. (2024). Research on the Application of Intelligent Technology in Document and Archives Management. *Journal of Computer Technology and Electronic Research, 1*, 44–58.

Hodel, T. (2022). Supervised and unsupervised: Approaches to machine learning for textual entities. In L. Jaillant (ed.), *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitised Archival Collections*. Bielefeld: transcript Verlag.

International Council on Archives. (1996). *ICA Code of Ethics*.

International Council on Archives. (2000). *ISAD(G): General International Standard Archival Description*.

International Council on Archives, Expert Group on Archival Description [ICA-EGAD]. (2023). *Records in Contexts – Conceptual Model, version 1.0*. Paris: International Council on Archives. Retrieved from https://www.ica.org/ica-network/expert-groups/egad/records-in-contexts-conceptual-model/ (accessed on 11.10.2026).

International Organization for Standardization (ISO). (2016). *ISO 15489-1:2016: Information and Documentation – Records Management. Part 1: Concepts and Principles*.

International Organization for Standardization (ISO). (2025). *ISO 14721:2025: Space data system practices – Reference model for an open archival information system (OAIS)*. Edition 3. Geneva: ISO.

InterPARES Trust AI. (2021). InterPARES Trust AI. Retrieved from https://interparestrustai.org/ (accessed on 12.8.2026).

Jaillant, L. (2022). How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions. *Archival Science, 22*, 417–436.

Jaillant, L. and Caputo, A. (2022). Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data. *AI & Society, 37*(3), 823–835.

Jaillant, L. and Rees, A. (2023). Applying AI to digital archives: trust, collaboration and shared professional ethics. *Digital Scholarship in the Humanities, 38*(2), 571–585.

Jaillant, L., Aske, K. and Caputo, A. (2025). The National Archives (UK). In L. Jaillant, C. Warwick, P. Gooding, K. Aske, G. Layne-Worthey and J. S. Downie (eds), *Navigating Artificial Intelligence for Cultural Heritage Organisations*. London: UCL Press. https://doi.org/10.14324/111.9781800088375

Li, J. and Wang, J. (2025). Intelligent Archive Management Based on Deep Learning Technology Driven by Artificial Intelligence. *IEEE Access, 13*, 42377–42387.

Lucas, N., Salah, C. and Vidal-Gorène, C. (2022). *New results for the text recognition of Arabic Maghribī manuscripts – managing an under-resourced script*. arXiv preprint arXiv:2211.16147.

Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M. and Conrad, M. (2018). Archival records and training in the age of big data. In J. Percell, L. C. Sarin et al. (eds), *Re-envisioning the MLS: Perspectives on the Future of Library and Information Science Education*. Bingley: Emerald.

Milovanović, M. (2025). Archival practice and artificial intelligence. *Atlanti*, 35(2), 45–58. https://doi.org/10.33700/2670-4579.35.2.45-58(2025)

Muehlberger, G., Seaward, L., Terras, M. et al. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5), 954–976.

Nair, A., Goh, Z. R., Liu, T. and Huang, A. Y. (2025). Web archives metadata generation with GPT-4o: Challenges and insights. *Information Technology and Libraries*, 44(2), 1–16.

National Archives and Records Administration [NARA]. (2025). *AI Compliance Plan* (18 September 2025). Retrieved from https://www.archives.gov/files/nara-2025-ai-compliance-plan-final-v1-09-18-2025.pdf (accessed on 11.10.2026).

National Archives and Records Administration [NARA]. (2026). *Guidance on Applying the Federal Records Act to Artificial Intelligence Materials* (memorandum, August 2026). Retrieved from https://www.archives.gov/files/records-mgmt/policy/nara-fra-ai-guidance.pdf (accessed on 11.10.2026).

National Archives of Australia. (n.d.). Information management for records created using Artificial Intelligence (AI) technologies. Retrieved from https://www.naa.gov.au/node/3660 (accessed on 11.10.2026).

National Institute of Standards and Technology [NIST]. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.AI.100-1

Nockels, J., Gooding, P., Ames, S. and Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Archival Science*, 22(3), 367–392. https://doi.org/10.1007/s10502-022-09397-0

Note, M. (n.d.). The Future of Archival Appraisal. Lucidea. Retrieved at https://lucidea.com/blog/archival-apprasial-and-ai/ (accessed on 12.8.2026).

Pignataro, T. C. and de Souza Neto, M. P. (2025). IA generativa na extração de metadados arquivísticos: um estudo baseado na norma ISAD(G). *Encontros Bibli*, 30, e103505. https://doi.org/10.5007/1518-2924.2025.e103505

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T. and Stuart, K. (2019). More human than human? Artificial intelligence in the archive. *Archives and Manuscripts*, 47(2), 179–203.

Saccucci, C. and Potter, A. (2025). Results of AI experimentation for cataloging at the Library of Congress. Paper presented at the 89th IFLA World Library and Information Congress, Astana, August 2025. Retrieved from https://repository.ifla.org/handle/20.500.14598/6861 (accessed on 11.10.2026).

Schwesig, R., Brich, I., Buder, J., Huff, M. and Said, N. (2023). Using artificial intelligence (AI)? Risk and opportunity perception of AI predict people's willingness to use AI. *Journal of Risk Research*, 26(10), 1053–1084.

Shinde, G., Kirstein, T., Ghosh, S. and Franks, P. (2025). Tracing the past, predicting the future: A systematic review of AI in archival science. *Proceedings of the Association for Information Science and Technology*, 62(1), 659–671. https://doi.org/10.1002/pra2.1286

Spina, S. (2023). *Artificial intelligence in archival and historical scholarship workflow: HTS and ChatGPT*. arXiv preprint arXiv:2308.02044.

SUMMARY

Archival institutions around the world face an unprecedented growth in the volume, variety and speed of the records they must appraise, describe, preserve and make accessible. Born-digital material – electronic mail, databases, social media, geospatial data and multimedia – is now created at a scale that traditional, labour-intensive archival methods cannot fully absorb, producing large backlogs of unprocessed and effectively inaccessible material. Artificial intelligence, encompassing machine learning, natural language processing, computer vision and, more recently, large language models, has become one of the principal responses to this challenge, and archival scholars and practitioners have experimented with AI-assisted transcription, automated subject indexing, sensitivity review, entity extraction and

predictive appraisal, generally reporting gains in processing speed and coverage alongside recurring concerns about accuracy, transparency and professional trust.

Despite the growing number of individual pilot applications, the scholarly and professional literature has offered comparatively little in the way of an integrated architectural vision showing how these separate AI tools could be combined into a single, coherent, standards-based archival system, and how such a system relates to established archival theory and professional ethics. This article addresses that gap in four steps: it reviews, in a structured narrative review of mainly 2019–2026 sources and a cross-study synthesis of eleven empirical evaluations, what is known about AI across archival functions; it proposes a conceptual five-layer reference architecture, the "intelligent archival ecosystem"; it tests the architecture against six publicly reported implementations and policy responses; and it specifies a protocol for validating the model empirically.

The article situates its proposal within core archival theory and standards, including continuum-based thinking about the record lifecycle, the third edition of the OAIS reference model (ISO 14721:2025), the Records in Contexts standard that now supersedes ISAD(G), and ISO 15489, and it draws on the trustworthiness principles of the multi-national InterPARES Trust AI project. The review shows that natural language processing and computer vision are the most mature technique families; that performance depends strongly on the interpretive demands of the task, with structured fields such as title and author reaching high accuracy in one national-library experiment while subject and genre remained below 50%; that generative models offer large cost savings but still produce inaccuracies and hallucinations that require human review; that recognition models for Arabic-script manuscripts typically need adaptation to each collection; that an experiment on AI-generated descriptions of photographs found that hands-on exposure to imperfect output lowered participants' trust and willingness to use AI; and that practitioner studies identify ethical, organisational and skills barriers alongside technical ones. Recent official guidance from national archives adds that AI-related materials – prompts, outputs, audit trails, models and training copies – may themselves be records that must be managed.

The five layers are: (1) ingestion and capture, where AI supports format identification, duplicate detection and optical character or handwriting recognition at

the point records enter the system; (2) AI-augmented description, classification and appraisal, where natural language processing and machine learning generate provisional, confidence-scored metadata and disposition recommendations for human review; (3) trust, provenance and authenticity, which maintains a verifiable record of which AI model produced or altered which piece of metadata and treats AI-related materials as records; (4) discovery and access, where AI supports semantic search and preliminary sensitivity classification while final release decisions remain a documented human responsibility; and (5) governance and human oversight, a cross-cutting layer comprising policy, role-based review workflows, explainability mechanisms and periodic bias auditing, which can be organised around the functions of the NIST AI Risk Management Framework.

The documentary analysis of six cases – handwritten text recognition with Transkribus, the Library of Congress's Exploring Computational Description experiment, GPT-4o metadata generation in Web Archive Singapore, AI projects at The National Archives (UK), machine-learning support for FOIA exemption review, and records-management guidance and governance instruments from the US and Australian national archives – supports the architecture's emphasis on human-in-the-loop design and shows that the published evidence is concentrated in capture, description and access, while provenance and governance remain the least evaluated layers. The analysis yields a design principle, field-level automation tiers, under which each metadata field and decision type is assigned a level of human review according to its measured accuracy against an institution-defined threshold.

The article then discusses the opportunities and risks associated with the architecture. Opportunities include the realistic prospect of addressing large descriptive and appraisal backlogs, more consistent application of descriptive standards, faster sensitivity review supporting timely release of government records, and the ability to surface thematic connections across dispersed holdings. Risks include algorithmic bias inherited from historical training data or from the archive's own past practices, the limited explainability of many machine-learning and large-language-model systems, the risk of factually inaccurate or fabricated descriptive content, and the loss of control over evidential AI materials held on third-party platforms. Drawing on studies of professional trust and of the compe-

tencies archivists need, the article argues that responsible adoption depends less on the sophistication of individual algorithms than on the governance, transparency and shared professional ethics surrounding their use, and that the archivist's role is best understood as that of orchestrator and auditor of a partially automated workflow who retains final authority over decisions of legal, ethical or historical significance.

The article concludes that the proliferation of discrete AI applications in archival practice can usefully be organised into a single, theory-grounded architecture designed from the outset around the archival principles of authenticity, reliability, integrity and usability. Because the proposed model is conceptual and based on literature synthesis and documentary analysis rather than empirical testing, validation is identified as the most pressing direction for future research: a multiple-case study of three to five institutions measuring performance, provenance completeness, governance coverage and the changing role of the archivist, together with continued collaboration between archival scholars, computer scientists and practitioners of the kind pursued by InterPARES Trust AI.

Typology: 1.03 Short Scientific Article