AI-Driven Enterprise Information Management and Intelligent Document Processing for Digital Transformation in Saudi Arabia under Vision 2030.

Abstract

Saudi organisations hold vast stores of scanned forms, contracts, correspondence and archival files, much of it in Arabic script, and Vision 2030 asks them to turn this material into reliable, reusable information for digital services. Artificial intelligence now offers the means to read, classify and extract content from such documents at scale, but recognition accuracy alone does not make information trustworthy, compliant or useful. This review examines how intelligent document processing (IDP) and enterprise information management (EIM) can be combined so that automated capture produces governed information assets that support digital transformation in the Kingdom. A structured evidence review of peer-reviewed journal studies from 2020 to 2025 was carried out, anchored by two seed surveys of Arabic optical character recognition and Arabic handwriting recognition, and closed at 30 sources. Evidence was charted against the document pipeline, the information lifecycle, Saudi regulatory anchors and transformation outcomes. The synthesis indicates that recognition of printed Arabic has matured on benchmark data, whereas handwriting, complex layouts, scarce domain-specific training data and post-recognition correction remain weak points. Layout-aware and multimodal models generalise better to unseen document formats, and large language models add flexible extraction but require human validation. On the governance side, records management, metadata, data quality, retention and preservation must be redesigned for machine participation, while Saudi studies point to data integration, inter-agency coordination and cybersecurity maturity as binding constraints. The paper proposes a governed document intelligence pipeline, a five-level maturity path and a control matrix linking each lifecycle stage to Saudi obligations, together with a research agenda.

UNDER PEER REVIEW IN IJAR

1. Introduction

Most organisational knowledge still arrives as documents. Invoices, permits, medical referrals, court filings, procurement files and personnel records are created, exchanged and stored long before they become rows in a database, and a large proportion of them exist only as images or loosely structured files. Enterprise information management (EIM) is the discipline that keeps such material authentic, findable, protected and disposed of on time, through records management, metadata, information governance and knowledge management. Intelligent document processing (IDP) is the set of artificial intelligence (AI) techniques that convert document images and files into structured, machine-usable content: optical character recognition (OCR) and handwriting recognition, layout analysis, document classification, information extraction and, more recently, document understanding with large language models (LLMs). The two fields have developed largely apart, one rooted in archival science and compliance, the other in computer vision and natural language processing.

Saudi Arabia makes their convergence urgent. Vision 2030 has placed digital government, data-driven services and private-sector productivity at the centre of national reform, and a qualitative case study of the Kingdom's digital government programme reports substantial progress in digitising public services alongside persistent difficulties in data integration, policy alignment and collaboration between agencies (Alkorbi & Alrwais, 2025). Those difficulties are, to a large extent, information management problems: services cannot be joined up if the documents and records beneath them are inconsistently captured, described and controlled. At the same time, the Arabic script poses specific technical obstacles. Its cursive, context-dependent letter forms, diacritics and ligatures make recognition harder than for Latin scripts, particularly for handwriting (Alrobah & Albahli, 2022; Faizullah et al., 2023).

A further reason for attention is regulatory. The Kingdom's personal data protection law, national cybersecurity controls issued by the National Cybersecurity Authority (NCA) and data-management standards issued by the National Data Management Office (NDMO) under the Saudi Data and Artificial Intelligence Authority place obligations on how documents containing personal or sensitive information are collected, processed, retained and shared. Automating document handling multiplies the speed and scale at which these obligations can be met or

UNDER PEER REVIEW IN IJAR

breached. An IDP deployment that extracts identity numbers from scanned applications, for instance, creates new copies of personal data, new derived records and new audit requirements.

Existing reviews address parts of this problem. Surveys of Arabic OCR and handwriting recognition concentrate on algorithms and datasets (Alrobah & Albahli, 2022; Faizullah et al., 2023; Najam & Faizullah, 2023); records-management scholarship examines the cultural and professional implications of AI without engaging the technical pipeline (Alaoui, 2025; Xie et al., 2022); and Saudi digital-government studies discuss adoption and coordination without examining document processing specifically (Alkorbi & Alrwais, 2025; Alshahrani, 2025). This review joins these strands. Its central argument is that IDP creates value for digital transformation only when its outputs are treated as governed information assets, with provenance, metadata, quality controls and retention rules attached at the moment of capture rather than added afterwards.

2. Objectives, Aim and Research Questions

The aim of the review is to build a framework, grounded in the reviewed evidence, that integrates AI-driven document processing with enterprise information management in Saudi organisations, and to identify the technical, governance and institutional conditions under which this integration yields measurable digital transformation outcomes. The review pursues four objectives. Objective 1 is to assess the state of the art in Arabic and bilingual document capture and understanding, covering printed and handwritten recognition, layout analysis, classification, field extraction, named-entity recognition and LLM-based understanding. Objective 2 is to examine how records management, metadata, data quality, retention, digital preservation and knowledge management are changing as AI enters the information lifecycle. The third is to map Saudi regulatory and institutional anchors, including personal data protection, NCA cybersecurity controls, NDMO data-management standards and digital-government coordination, onto practical controls. The fourth is to derive a framework, a maturity path and an agenda for research and practice.

Three research questions guide the analysis. RQ1: How mature are AI techniques for capturing and understanding Arabic-script and bilingual enterprise documents, and at which stages of the pipeline do they remain unreliable? RQ2: What changes to information governance practice are needed when documents are captured, described and interpreted by machines rather than by

UNDER PEER REVIEW IN IJAP

people? RQ3: How can Saudi organisations connect document intelligence and information management to digital transformation outcomes while meeting national data-protection and cybersecurity obligations?

Several terms are used in a specific sense. A document is any recorded unit of information with a fixed form, whether born-digital or scanned. A record is a document retained as evidence of a business activity, with requirements for authenticity, integrity and retention. Information governance is the set of policies, roles and controls through which an organisation decides how information is created, used, protected and disposed of. Document intelligence refers collectively to the IDP techniques listed above. Digital transformation outcomes include faster and more accurate service delivery, lower processing cost, better use of information in decisions and stronger compliance and trust.

3. Review Methodology

A structured evidence review was chosen because the relevant literature spans computer science benchmarks, records and archival studies, information-systems research and legal analysis. A design that charts evidence against an explicit framework, rather than pooling effect sizes, suits this heterogeneity. The protocol defined, before searching, the questions, eligibility rules, the charting framework and the order of screening, and each decision was logged so that the corpus can be reconstructed.

Two seed studies anchored the review. Faizullah et al. (2023) survey Arabic OCR across the full recognition process and identify research gaps, and Alrobah and Albahli (2022) classify deep-learning approaches to offline Arabic handwriting recognition and discuss the challenges specific to the script. The handwriting survey was produced at a Saudi university, and both seeds bear directly on the capture stage of IDP. Their reference lists and citing articles were examined to identify technical work within the review window. Because the seeds are technical, separate searches were needed for information governance, records management and Saudi institutional evidence.

Searches were run against a multidisciplinary bibliographic metadata service for journal articles dated from January 2020 to December 2025. Search strings combined capture and understanding terms ("optical character recognition", "handwriting recognition", "layout analysis", "document

UNDER PEER REVIEW IN IJAP

understanding", "information extraction", "named entity recognition", "large language model", invoice), information management terms ("records management", "information governance", metadata, "data quality", "digital preservation", "knowledge management", "data maturity"), and context terms (Arabic, "Saudi Arabia", "Vision 2030", "digital government", cybersecurity, "data protection"). Capture and information management strings were run both with and without context terms, so that strong international evidence could be retained where Saudi evidence was absent.

Inclusion required a peer-reviewed journal article whose issue year, or online year when no issue had been assigned, fell between 2020 and 2025; a resolvable digital object identifier with complete bibliographic metadata; and a substantive bearing on one or more of the four objectives. Records were excluded when they were conference papers, book chapters, preprints or editorials; when their document-processing focus lay in a domain without enterprise relevance, such as scene text or sign language; when they concerned clinical or financial prediction rather than documents or information management; or when their journal metadata lacked volume or page information. Candidates whose final issue was dated 2026 were removed even when they had appeared online earlier.

Screening proceeded in three passes: titles, then abstracts, then the full text or, where the full text was inaccessible, the abstract together with publisher landing-page information. In the final pass, each candidate was judged on relevance to the research questions, clarity of method, adequacy of evidence for its claims and transferability to Saudi organisations. Benchmark studies were read for what their datasets and metrics can and cannot show; conceptual papers were treated as sources of propositions rather than evidence of effects.

The final corpus comprises 30 journal articles. Eight are surveys or literature reviews, ten are technical studies that propose and evaluate methods or datasets, seven are conceptual or framework papers, and five are empirical organisational, case-study or legal analyses. Seven sources examine Saudi institutions directly, and eight address Arabic-script processing. Each source was charted against four dimensions: the pipeline stage (capture, layout, classification, extraction, understanding, validation); the information lifecycle stage (creation, description, use, retention, preservation, disposal); the Saudi anchor involved (data protection, cybersecurity, data management, digital government); and the outcome reported or implied. Synthesis moved from

UNDER PEER REVIEW IN IJAR

summaries within each dimension, through comparison across dimensions, to construction of the framework, maturity path and control matrix.

4. Capturing and Understanding Arabic-Script Documents

Recognition is the foundation of every document pipeline, and for Arabic it remains the stage where most risk originates. Faizullah et al. (2023) describe the complete OCR process, from pre-processing and segmentation to feature extraction, recognition and post-processing, and note that each step faces script-specific problems, including connected letters, position-dependent shapes and dots that distinguish otherwise identical characters. Alrobah and Albahli (2022) show that deep learning has become the dominant approach to offline handwritten Arabic, but they also identify unresolved difficulties in segmentation, dataset size and variability between writers. Najam and Faizullah (2023), reviewing studies from 2020 to 2023, conclude that convolutional and recurrent networks trained with connectionist temporal classification are the most suitable architecture for Arabic handwritten OCR, because they handle long textual dependencies with less complexity than transformer or generative adversarial alternatives. They also find that post-OCR text correction is under-researched and that high-quality, domain-specific handwritten datasets are lacking.

Benchmark results for printed text are now very strong. Mosbah et al. (2024) report a character error rate of 0.01% on the P-KHATT line images and 0.03% on the APTI word images, and a word error rate of 1.09% on the IFN/ENIT handwritten word dataset, using a network that combines convolutional and bidirectional recurrent layers. Earlier non-deep approaches illustrate how far the field has moved: a printed Arabic OCR system based on thinning, chain-code features and compression-based classification reached 77.3% accuracy on paragraph images (Al Ghamdi, 2022). Benchmark accuracy should nevertheless be interpreted with care. Datasets of isolated words or clean lines do not reproduce the stamps, signatures, tables, mixed Arabic–English text and degraded scans that dominate enterprise archives, so a figure achieved on a benchmark is an upper bound rather than an operational expectation.

The enterprise case for recognition is broader than data entry. Faizullah et al. (2023) observe that automatic text extraction improves productivity and accessibility and supports the preservation of historical documents, three purposes that map onto distinct information management goals: operational efficiency, inclusive service delivery and long-term memory. Each purpose implies a

UNDER PEER REVIEW IN IJAR

different tolerance for error. A search index for an archive can tolerate occasional misread words, whereas a field that populates a payment or a legal decision cannot. Recognition quality should therefore be specified per use rather than per system, and the same scanned file may legitimately pass through different validation routes depending on where its content will be used. This principle, simple as it is, is often absent from procurement specifications that request a single headline accuracy figure.

Layout analysis determines whether recognised text can be assigned to the right field, paragraph or table cell. Elanwar et al. (2021) released BE-Arabic-9K, more than 9,000 scanned pages from over 700 Arabic books, 1,500 of which were manually segmented, and their fine-tuned region detector reached 99.4% accuracy in separating text from non-text blocks. The scarcity of such annotated resources is a recurring theme. Banerjee et al. (2024) propose semi-supervised layout analysis that learns from a limited number of annotated categories and generalises to new ones, while De Nardin et al. (2025) find that combining cross-domain and in-domain transfer learning yields the best layout segmentation of handwritten documents and speeds up model convergence. For Saudi organisations, these results suggest that sizeable gains are possible without labelling every document type, provided that a representative sample of local layouts is annotated.

Classification and extraction convert pages into business data. Multimodal pre-training that combines text, layout and image features has become the leading approach: GraphDoc, pre-trained on 320,000 unlabelled documents with a graph attention mechanism that links each text block to its neighbours, achieved state-of-the-art results on public document-understanding benchmarks (Z. Zhang et al., 2023). Invoice studies show why layout awareness matters in enterprise settings. Krieger et al. (2023) found that, when training data are skewed towards a few frequent suppliers, the accuracy gap between familiar and unfamiliar layouts was much larger for grid-based and random-forest models than for a layout-aware transformer, which also performed best overall. Ha and Horák (2022) combined OCR output with text and layout features to recover invoice data in 90% of English and 88% of Czech cases, and Arslan et al. (2024) used synthetic invoice generation with automatic labelling, improving accuracy by 8.4% over baseline models when the detector was fine-tuned on new templates. Synthetic data and careful evaluation design are therefore as important as model choice.

UNDER PEER REVIEW IN IJAR

Named-entity recognition (NER) extends extraction from fixed fields to names, organisations, places, dates and amounts embedded in free text. Moussaoui and Loqman (2024) trace Arabic NER from rule-based through statistical to deep-learning methods and catalogue the datasets, gazetteers and annotation schemes available. Al-Duwais et al. (2024) benchmarked multilingual language models for Arabic cross-lingual NER and found that GigaBERT performed best and that language-adaptive pre-training was the most effective adaptation method, underlining the value of Arabic-specific knowledge when transferring from English. Large language models add another layer. Kalyan (2024) surveys the GPT-3 family of models and their abilities in downstream tasks, data labelling and data augmentation, while also discussing their robustness. For IDP, these abilities suggest a role in extracting information from document types with no training data and in generating labelled examples, but they do not remove the need for verification.

Human validation is the final stage of a dependable pipeline. Wu et al. (2022) classify human-in-the-loop machine learning into work that improves data processing, work that intervenes in model training and system designs that integrate human contributions independently of the model. Applied to IDP, the three categories correspond to human correction of recognition and extraction errors, active selection of uncertain documents for labelling, and workflow designs that route low-confidence fields to reviewers. Table 1 consolidates the evidence across the pipeline, recording the Arabic-specific challenges at each stage and the quality measures that organisations should monitor.

Table 1. Evidence on the document intelligence pipeline for Arabic-script and bilingual enterprise documents

Pipeline stage	Representative evidence	Arabic-specific or enterprise challenge	Quality measures to monitor	Principal sources
Recognition (printed and handwritten)	Near-perfect character error rates on printed benchmarks; deep sequence models preferred for handwriting	Cursive, position-dependent letters and dots; writer variability; scarce domain-specific handwritten data; weak post-OCR correction	Character and word error rates on in-house samples; correction rate after review	Alrobah and Albahli (2022); Faizullah et al. (2023); Mosbah et al. (2024); Najam and Faizullah (2023)

UNDER PEER REVIEW IN IJAR

Pipeline stage	Representative evidence	Arabic-specific or enterprise challenge	Quality measures to monitor	Principal sources
Layout analysis	High text versus non-text accuracy on Arabic book pages; semi-supervised and transfer learning reduce annotation needs	Complex pages with tables, stamps and mixed scripts; limited annotated Arabic layouts	Region detection accuracy; reading-order errors	Banerjee et al. (2024); De Nardin et al. (2025); Elanwar et al. (2021)
Classification and field extraction	Multimodal and layout-aware models generalise better to unseen layouts; synthetic data aids new templates	Skewed supplier or form populations; frequent template change	Field-level accuracy on unseen layouts; straight-through rate	Arslan et al. (2024); Ha and Horák (2022); Krieger et al. (2023); Z. Zhang et al. (2023)
Entity recognition and LLM-based understanding	Arabic-adapted pre-training improves cross-lingual NER; LLMs support labelling and flexible extraction	Dialect and domain vocabulary; unverified or fabricated outputs	Entity-level F1; rate of unsupported values found in review	Al-Duwais et al. (2024); Kalyan (2024); Moussaoui and Loqman (2024)
Human validation	Human input improves data, training and system design	Reviewer workload; consistency of corrections	Share of fields reviewed; reviewer agreement; turnaround	Wu et al. (2022)

5. From Documents to Governed Information Assets

Extracting text and fields is not the same as managing information. Figure 1 presents the governed document intelligence pipeline proposed in this review. Documents enter through capture channels, pass through image enhancement and recognition, layout analysis, classification and extraction, and LLM-assisted understanding, and then reach a validation stage in which human reviewers confirm low-confidence results. Validated outputs flow into three destinations: the records repository, where authoritative copies are kept with retention rules; operational systems that use the extracted data; and a knowledge layer that supports search and reuse. Beneath the pipeline runs a governance plane that attaches provenance, metadata, access classification and retention schedules at every stage, together with the controls required for data protection and cybersecurity.

UNDER PEER REVIEW IN IJAR

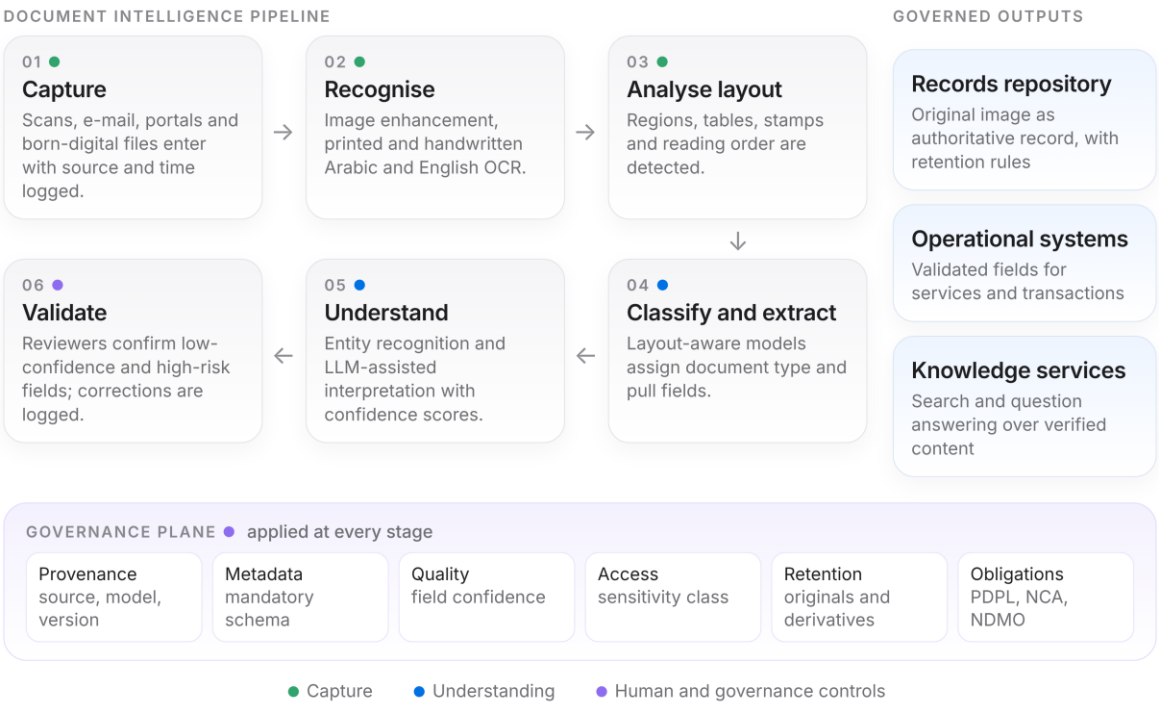


Figure 1. Governed document intelligence pipeline linking capture, understanding and validation to a governance plane.

Records management is the first discipline that must adapt. Alaoui (2025), drawing on an information culture framework, argues that integrating AI into records management depends on visible aspects of culture, such as technology use, collaboration and horizontal communication; on policies that define the purposes of automation and the roles of those involved; and on individual perceptions that shape whether staff trust AI tools. Records managers, in this view, take on new roles in change management and in building information competencies. Xie et al. (2022), analysing professional competencies against indicators of susceptibility to AI, conclude that records and information management is currently resistant to AI but not immune to it, and that some established principles are not yet ready to accommodate machine judgement. Shonhe (2025) proposes a conceptual framework for AI readiness and adoption in records management that combines technology acceptance, technology readiness and cognitive appraisal theories, while noting that empirical tests are still lacking. Together, these studies imply that IDP projects in Saudi organisations should involve records professionals from the outset rather than presenting them with automated outputs after deployment.

UNDER PEER REVIEW

Metadata and data quality determine whether captured information remains usable. Miller et al. (2024) document a lack of standard terminology for data quality and propose a data model that extends the dimensions of the ISO 25012 standard with four more: governance, usefulness, quantity and semantics. These additions are directly relevant to IDP, because a machine-extracted value may be accurate at the character level yet lack the provenance, context or meaning needed for a business decision. Saudi evidence shows that quality problems persist even in mature national initiatives: Alogaiel and Alrwais (2023), assessing the Saudi open government data programme launched in 2011, identified quality shortcomings that limit reuse and proposed a weighted assessment framework compatible with the national open data portal. Organisations should therefore define quality measures for extracted data, such as field-level confidence, completeness against mandatory schema elements and agreement with source systems, and report them alongside recognition accuracy.

Retention and preservation raise a different set of questions. Automated pipelines create derivative objects, including recognised text, extracted fields, embeddings and model logs, each of which may need its own retention rule. Teel (2024) asks whether AI could change how preservation is practised, since traditional methods are costly and labour-intensive, and suggests that AI may reshape the preservation of historical collections. In the Saudi context, where ministries and companies hold long runs of paper files in Arabic, the combination of high-accuracy layout analysis and recognition with archival metadata standards could make such collections searchable for the first time. Preservation, however, demands that the original image remain the authoritative record and that recognised text be labelled as a derived, potentially imperfect representation.

Knowledge management is the point at which document intelligence becomes organisational learning. Q. Zhang et al. (2025) develop a framework in which generative AI supports four knowledge processes (creating knowledge, storing and retrieving it, sharing and transferring it, and applying it) and they argue that human–AI collaboration is needed to counter data bias, information overload and integration problems. Applied to EIM, the framework suggests that validated document content should feed enterprise search and question-answering services, but that the organisation must retain clear lines between authoritative records and AI-generated

UNDER PEER REVIEW IN IJAR

summaries. Without such lines, a knowledge service built on LLMs could present unverified extractions as fact, undermining the trust that records management exists to protect.

6. Saudi Regulatory Anchors, Transformation Outcomes and a Maturity Path

Saudi organisations deploy document intelligence within a dense and evolving regulatory environment. J. Sarabdeen and Mohamed Ishak (2025), testing Saudi and Malaysian health-data legislation against what the European Union demands of third countries under its General Data Protection Regulation, report that Saudi health data were protected by general laws until the newly passed data protection law was implemented in 2024. For IDP, the practical consequence is that every extraction of personal data from a document must have a lawful basis, a defined purpose and a retention limit, and that derived copies fall within the same obligations as the source. Cybersecurity requirements apply in parallel. Almomani et al. (2021) note that Saudi frameworks, including the NCA Essential Cybersecurity Controls, lacked a practical mechanism for continuous assessment, and they propose a maturity assessment framework that maps the regulations Saudi organisations must comply with onto measurable levels. Alhalafi and Veeraraghavan (2021), comparing the Saudi cybersecurity policy framework with those of the United States, India and Singapore, recommend updating cybercrime legislation and adopting a master cybersecurity plan. Document pipelines, which concentrate sensitive information in scanning queues, model-training sets and extraction logs, are exactly the kind of asset these controls are meant to protect.

Institutional coordination is the second anchor. Alkorbi and Alrwais (2025) identify institutional coordination, mechanisms for monitoring progress and sustained spending on digital skills and infrastructure as essentials of Saudi digital government transformation, and Alshahrani (2025) finds that legacy-system integration, interoperability, financial limits and organisational resistance hinder the adoption of emerging technologies across six government sectors. Okuyucu and Yavuz (2020) distinguish big-data maturity models that assess organisations from those that assess states, the latter emphasising legislative frameworks and national policy. This distinction is useful for Saudi Arabia, where organisational information maturity depends partly on national enablers such as common data standards and shared platforms. Evidence from the private sector links these capabilities to performance: in a survey of 164 executives and IT managers in Saudi

UNDER PEER REVIEW IN IJAR

service firms, digital transformation, customer experience and IT innovation each improved firm performance, with customer experience showing the strongest effect (Masoud & Basahel, 2023).

Data sharing between organisations deserves particular attention, because it is where the value of extracted information multiplies and where governance failures propagate. The assessment framework proposed by Alogaiel and Alrwais (2023) for Saudi open government data weighs quality characteristics that determine whether published datasets can be reused, and its logic extends naturally to data derived from documents. Before extracted fields are passed to another agency, a shared service or an analytics platform, they should carry quality scores, provenance and a statement of the purpose for which they were collected. Where documents contain personal data, sharing must also respect the purpose limitation and storage rules of the data protection regime (J. Sarabdeen & Mohamed Ishak, 2025). In practice, this means that the metadata schema of an IDP platform is as consequential for digital government as its recognition engine.

Table 2 translates these anchors into a control matrix. For each stage of the information lifecycle it identifies the principal Saudi obligation, the AI-specific control required and the accountable role. The matrix makes clear that governance cannot be applied at the end of the pipeline: data minimisation belongs at capture, access classification at classification, purpose limitation at extraction, provenance at validation, and retention and secure disposal at the repository.

Table 2. Control matrix linking the information lifecycle to Saudi obligations and AI-specific controls

Lifecycle stage	Principal Saudi anchor	AI-specific control	Accountable role
Capture and intake	Personal data protection law: lawful basis and minimisation	Capture only required pages and fields; mask unneeded identifiers; log source and time of capture	Process owner with data protection officer
Recognition and classification	NCA cybersecurity controls: protection of processing environments	Secure, access-controlled processing queues; classification of document sensitivity on arrival	Cybersecurity team
Extraction and understanding	Data protection: purpose limitation; NDMO data-management standards	Extract only fields tied to a documented purpose; record model version and confidence for each value	Data steward
Validation	Records authenticity and integrity requirements	Human review of low-confidence and high-risk fields; audit trail of corrections	Records manager with business reviewers
Storage, use and sharing	Digital-government interoperability and data-sharing rules	Mandatory metadata schema; quality checks before release to other systems or agencies	Information governance lead

UNDER PEER REVIEW IN IJAP

Lifecycle stage	Principal Saudi anchor	AI-specific control	Accountable role
Retention, preservation and disposal	Retention schedules; data protection storage limitation	Retain original image as authoritative record; apply retention to derived text, fields and training sets; verified disposal	Records manager

Figure 2 sets out a five-level maturity path. At Level 1, digitised, documents are scanned and stored as images with minimal metadata. Level 2, indexed, adds OCR, which produces searchable text and basic document types are assigned. Level 3, understood, is reached when layout-aware extraction and NER produce structured fields with confidence scores, and low-confidence items are routed to reviewers. In Level 4, governed, provenance, quality metrics, retention rules and data-protection controls are attached automatically and audited. Finally, in Level 5, knowledge-driven, validated information feeds cross-functional services, enterprise search and decision support, and performance is tracked against transformation outcomes. The levels are cumulative. An organisation that jumps to LLM-based knowledge services at Level 5 without the controls of Level 4 gains speed at the cost of accountability, and an organisation that stops at Level 2 has digitised paper without transforming its processes.

Progression through the levels should be measured with indicators drawn from both disciplines. Capture indicators include character and word error rates on samples of the organisation's own documents, field-level extraction accuracy and the share of documents processed without human intervention. Governance indicators include the proportion of records with complete mandatory metadata, the share of extracted personal data covered by a documented purpose and retention rule, and findings from cybersecurity assessments. Outcome indicators include processing time per document, service turnaround for citizens or customers, reuse of information across units and user satisfaction. Reporting these indicators together prevents the common pattern in which technical accuracy improves while governance debt accumulates.

UNDER PEER REVIEW IN IJAR

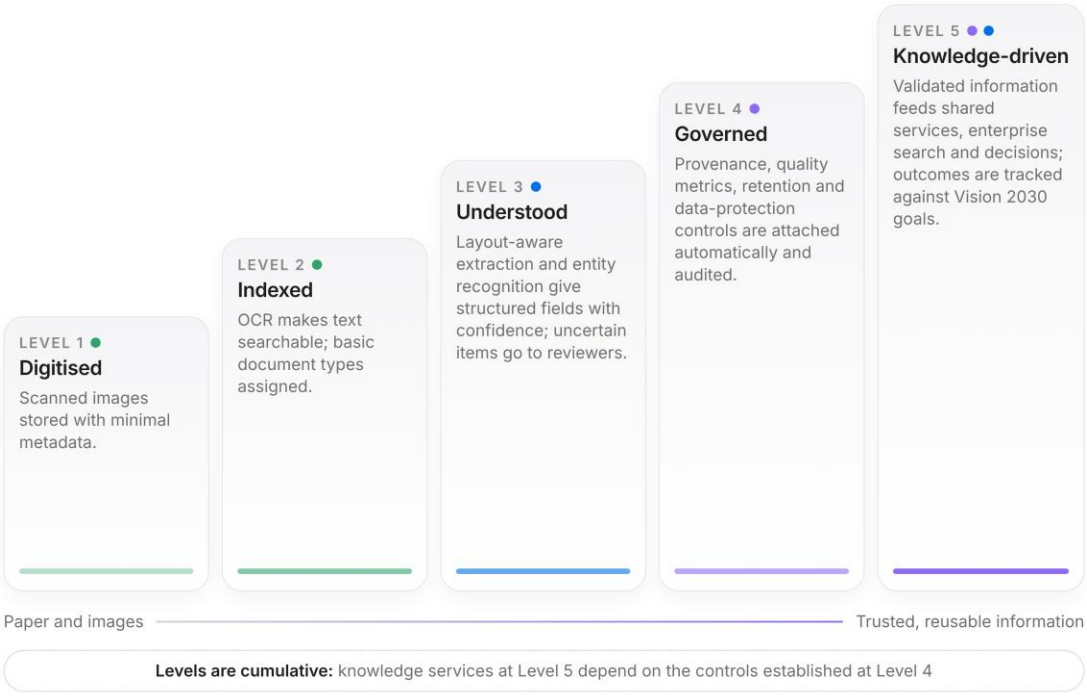


Figure 2. Five-level maturity path for AI-driven enterprise information management.

7. Discussion, Research Agenda and Limitations

In answer to RQ1, the evidence indicates that AI capture of printed Arabic text has reached high accuracy on benchmark data, and that layout-aware, multimodal and transfer-learning methods are reducing the dependence on large annotated corpora. Handwriting, degraded historical material, mixed-script documents and post-recognition correction remain weak, and the scarcity of domain-specific Arabic datasets is the most frequently cited constraint. LLMs extend extraction to unseen document types but introduce verification needs of their own. In answer to RQ2, information governance must move from periodic, manual control to rules embedded in the pipeline: metadata, provenance, quality measures and retention must be generated at capture, and records professionals need new roles in designing and auditing automated processes. In answer to RQ3, Saudi organisations can link document intelligence to transformation outcomes by treating extraction outputs as governed assets, by aligning controls with data-protection and cybersecurity obligations from the outset, and by measuring progress on a maturity path that includes service outcomes.

UNDER PEER REVIEW

Three propositions follow from the synthesis. First, the operational accuracy of Arabic IDP depends more on the representativeness of local training and evaluation data than on the choice among current model architectures. Second, organisations that attach metadata, provenance and retention rules at the point of capture will incur lower compliance costs and achieve greater reuse of information than those that apply governance retrospectively. Third, the contribution of document intelligence to digital transformation in Saudi organisations is mediated by inter-agency and cross-functional data integration, so that IDP investments without integration will yield local efficiency but limited service improvement.

The research agenda is correspondingly broad. Shared, privacy-protected Arabic datasets of administrative documents, including handwritten forms, stamps and bilingual layouts, would allow results to be compared under realistic conditions. Evaluation of LLM-based extraction on Arabic enterprise documents, including rates of fabricated or misattributed values, is largely missing. Empirical studies of AI readiness in Saudi records and information management units would test the conceptual frameworks now available. Legal research is needed on the status of machine-extracted data and AI-generated summaries as records, and on how retention and disposal obligations apply to model training sets. Finally, longitudinal studies linking maturity levels to service and financial outcomes would show whether the governed pipeline delivers the benefits argued here.

Practitioners can draw four immediate lessons. Before procuring IDP platforms, organisations should benchmark candidate systems on samples of their own documents rather than relying on published accuracy. Records managers, data protection officers and cybersecurity teams should co-own pipeline design with technical teams. Confidence thresholds and review queues should be set according to the consequences of an error, with stricter thresholds for identity, financial and legal fields. Finally, original images should be preserved as the authoritative record, and derived text should be labelled with its provenance and confidence.

The review has limitations. The corpus is restricted to 30 journal articles, and the exclusion of conference proceedings omits a substantial part of the document-analysis literature, where many advances first appear. Searches relied on one metadata service and on the citation trails of two seeds, and studies published in Arabic were not included. A single reviewer performed screening and charting. Several technical studies report results on public benchmarks rather than Saudi

UNDER PEER REVIEW IN IJAR

enterprise documents, and several governance sources are conceptual. The framework, maturity path and control matrix are proposals derived from synthesis and require testing in Saudi organisations.

8. Conclusion

This review set out to explain how AI-driven document processing and enterprise information management can be combined to support digital transformation in Saudi Arabia. The evidence shows that the techniques for reading Arabic-script documents have advanced quickly, especially for printed text and layout analysis, while handwriting, domain-specific data and verification remain open problems. It also shows that technical capture is only half of the task. Information becomes a dependable asset for digital services only when metadata, provenance, quality measures, retention and security are built into the pipeline and when records professionals help to design it.

The governed document intelligence pipeline, the five-level maturity path and the control matrix offer Saudi organisations a way to sequence investment and to align it with national data-protection, cybersecurity and data-management obligations. Their common message is that the value of document intelligence lies not in the number of pages converted but in the trustworthy, reusable information that results. Organisations that govern information at the moment of capture will be best placed to turn their document archives into a foundation for the services envisaged by Vision 2030.

UNDER PEER REVIEW IN IJAR

References

1. Al Ghamdi, M. A. (2022). A novel approach to printed Arabic optical character recognition. *Arabian Journal for Science and Engineering*, *47*(2), 2219–2237. https://doi.org/10.1007/s13369-021-06163-9

2. Alaoui, S. (2025). Artificial intelligence and records management in contemporary organizations: What cultural aspects are required? Insights from the information culture framework (ICF). *Records Management Journal*, *35*(2), 109–117. https://doi.org/10.1108/rmj-08-2023-0041

3. Al-Duwais, M., Al-Khalifa, H., & Al-Salman, A. (2024). A benchmark evaluation of multilingual large language models for Arabic cross-lingual named-entity recognition. *Electronics*, *13*(17), Article 3574. https://doi.org/10.3390/electronics13173574

4. Alhalafi, N., & Veeraraghavan, P. (2021). Cybersecurity policy framework in Saudi Arabia: Literature review. *Frontiers in Computer Science*, *3*, Article 736874. https://doi.org/10.3389/fcomp.2021.736874

5. Alkorbi, S., & Alrwais, O. (2025). An empirical case study of digital government transformation in Saudi Arabia. *Information*, *16*(12), Article 1110. https://doi.org/10.3390/info16121110

6. Almomani, I., Ahmed, M., & Maglaras, L. (2021). Cybersecurity maturity assessment framework for higher education institutions in Saudi Arabia. *PeerJ Computer Science*, *7*, Article e703. https://doi.org/10.7717/peerj-cs.703

7. Alogaiel, N. F., & Alrwais, O. A. (2023). An assessment of the quality of open government data in Saudi Arabia. *IEEE Access*, *11*, 61560–61599. https://doi.org/10.1109/access.2023.3285611

8. Alrobah, N., & Albahli, S. (2022). Arabic handwritten recognition using deep learning: A survey. *Arabian Journal for Science and Engineering*, *47*(8), 9943–9963. https://doi.org/10.1007/s13369-021-06363-3

9. Alshahrani, A. (2025). Adopting emerging technologies in digital government: A multi-case analysis of drivers, enablers, and challenges in Saudi Arabia. *Digital Government: Research and Practice*, *6*(3), 1–26. https://doi.org/10.1145/3719297

UNDER PEER REVIEW IN IJAR

10. Arslan, H., Işık, Y. E., & Görmez, Y. (2024). A deep learning-based solution for digitization of invoice images with automatic invoice generation and labelling. *International Journal on Document Analysis and Recognition (IJDAR)*, *27*(1), 97–109. https://doi.org/10.1007/s10032-023-00449-4

11. Banerjee, A., Biswas, S., Lladós, J., & Pal, U. (2024). SemiDocSeg: Harnessing semi-supervised learning for document layout analysis. *International Journal on Document Analysis and Recognition (IJDAR)*, *27*(3), 317–334. https://doi.org/10.1007/s10032-024-00473-y

12. De Nardin, A., Zottin, S., Piciarelli, C., Foresti, G. L., & Colombi, E. (2025). In-domain versus out-of-domain transfer learning for document layout analysis. *International Journal on Document Analysis and Recognition (IJDAR)*, *28*(2), 161–175. https://doi.org/10.1007/s10032-024-00497-4

13. Elanwar, R., Qin, W., Betke, M., & Wijaya, D. (2021). Extracting text from scanned Arabic books: A large-scale benchmark dataset and a fine-tuned Faster-R-CNN model. *International Journal on Document Analysis and Recognition (IJDAR)*, *24*(4), 349–362. https://doi.org/10.1007/s10032-021-00382-4

14. Faizullah, S., Ayub, M. S., Hussain, S., & Khan, M. A. (2023). A survey of OCR in Arabic language: Applications, techniques, and challenges. *Applied Sciences*, *13*(7), Article 4584. https://doi.org/10.3390/app13074584

15. Ha, H., & Horák, A. (2022). Information extraction from scanned invoice images using text analysis and layout features. *Signal Processing: Image Communication*, *102*, Article 116601. https://doi.org/10.1016/j.image.2021.116601

16. Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, Article 100048. https://doi.org/10.1016/j.nlp.2023.100048

17. Krieger, F., Drews, P., & Funk, B. (2023). Automated invoice processing: Machine learning-based information extraction for long tail suppliers. *Intelligent Systems with Applications*, *20*, Article 200285. https://doi.org/10.1016/j.iswa.2023.200285

UNDER PEER REVIEW IN IJAR

18. Masoud, R., & Basahel, S. (2023). The effects of digital transformation on firm performance: The role of customer experience and IT innovation. *Digital*, *3*(2), 109–126. https://doi.org/10.3390/digital3020008

19. Miller, R., Whelan, H., Chrubasik, M., Whittaker, D., Duncan, P., & Gregório, J. (2024). A framework for current and new data quality dimensions: An overview. *Data*, *9*(12), Article 151. https://doi.org/10.3390/data9120151

20. Mosbah, L., Moalla, I., Hamdani, T. M., Neji, B., Beyrouthy, T., & Alimi, A. M. (2024). ADOCRNet: A deep learning OCR for Arabic documents recognition. *IEEE Access*, *12*, 55620–55631. https://doi.org/10.1109/access.2024.3379530

21. Moussaoui, T. E., & Loqman, C. (2024). Advancements in Arabic named entity recognition: A comprehensive review. *IEEE Access*, *12*, 180238–180266. https://doi.org/10.1109/access.2024.3491897

22. Najam, R., & Faizullah, S. (2023). Analysis of recent deep learning techniques for Arabic handwritten-text OCR and post-OCR correction. *Applied Sciences*, *13*(13), Article 7568. https://doi.org/10.3390/app13137568

23. Okuyucu, A., & Yavuz, N. (2020). Big data maturity models for the public sector: A review of state and organizational level models. *Transforming Government: People, Process and Policy*, *14*(4), 681–699. https://doi.org/10.1108/tg-09-2019-0085

24. Sarabdeen, J., & Mohamed Ishak, M. M. (2025). A comparative analysis: Health data protection laws in Malaysia, Saudi Arabia and EU General Data Protection Regulation (GDPR). *International Journal of Law and Management*, *67*(1), 99–119. https://doi.org/10.1108/ijlma-01-2024-0025

25. Shonhe, L. (2025). Conceptual framework to explore artificial intelligence technology (AIT) readiness and adoption intention in records and information management (RIM) practices: A proposal. *Records Management Journal*, *35*(1), 18–34. https://doi.org/10.1108/rmj-09-2023-0046

26. Teel, Z. (2024). Artificial intelligence's role in digitally preserving historic archives. *Preservation, Digital Technology & Culture*, *53*(1), 29–33. https://doi.org/10.1515/pdtc-2023-0050

UNDER PEER REVIEW IN IJAR

27. Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

28. Xie, S. L., Siyi, L., & Han, R. (2022). Competing with artificial intelligence – can the records and information management profession withstand the challenge? *Records Management Journal*, *32*(2), 151–169. https://doi.org/10.1108/rmj-08-2021-0033

29. Zhang, Q., Zuo, J., & Yang, S. (2025). Research on the impact of generative artificial intelligence (GenAI) on enterprise innovation performance: A knowledge management perspective. *Journal of Knowledge Management*, *29*(7), 2238–2257. https://doi.org/10.1108/jkm-10-2024-1198

30. Zhang, Z., Ma, J., Du, J., Wang, L., & Zhang, J. (2023). Multimodal pre-training based on graph attention network for document understanding. *IEEE Transactions on Multimedia*, *25*, 6743–6755. https://doi.org/10.1109/tmm.2022.3214102

UNDER PEER REVIEW IN IJAR