End-to-end digitisation workflows with intelligent document processing for university theses and historical newspapers

Niklas Zimmer
FCTEC
niklas@fctec.co.za
https://orcid.org/0000-0001-8078-0403

Received: 02 June 2025
Revised: 25 October 2025
Accepted: 07 November 2025

Abstract

Archives and records management are entering an era defined by artificial intelligence, requiring digitisation workflows that are both efficient and contextually appropriate. This study employed a qualitative multiple-case study design to examine two end-to-end digitisation workflows that integrate archival-grade overhead and feed scanners with intelligent document processing (IDP) systems for automated data extraction, validation, and export. The two case studies focused, respectively, on university theses and historical newspapers, enabling a comparative evaluation of workflow efficiency, metadata accuracy, and system interoperability. The study found that each workflow conformed to the international imaging benchmark ISO 19264-1: 2021 – Photography – Archiving systems – Image quality analysis – Part 1: Reflective originals, ensuring measurable image quality control. A hybrid scanning strategy combining InoTec SCAMAX 6-series and Zeutschel OS A-series scanners was supported by a locally hosted IDP platform, NEURO.AI. The findings demonstrated measurable improvements in metadata accuracy with confidence levels exceeding 90% throughput, and repository interoperability compared to manual indexing. The results confirmed that integrated, intelligent digitisation workflows can substantially improve preservation quality and operational efficiency, offering a scalable, standards-aligned model for digital transformation in African archival and library institutions.

Keywords: End-to-end digitisation, workflows, intelligent document processing, university theses, historical newspapers

1. Introduction and background to the study

The archival and records management professions are undergoing a profound transformation due to the integration of artificial intelligence (AI) and digitisation technologies. These shifts are not merely technical; they represent a structural redefinition of how heritage institutions manage, preserve, and provide access to information resources. Digitisation is increasingly recognised not as a stand-alone activity but as a strategic enabler for intelligent archival management, metadata enrichment, and scalable access (Gilliland 2014).

In the African context, particularly in South Africa, public archives and libraries continue to face persistent challenges, including ageing physical collections, under-resourced infrastructures, and fragmented digital strategies (Ngoepe 2022). These systemic constraints continue to limit large-scale digital preservation and access initiatives. Nevertheless, there is

Licenced under Creative Commons: CC BY-NC-SA
https://dx.doi.org/10.4314/jsasa.v58i1.7
ISSN: 1012-2796
©SASA 2025

growing policy-level and institutional recognition of the need to transition towards standards-based, technology-enabled archival practices (IFLA 2020; UNESCO 2023). Recent regional and professional initiatives have highlighted the potential of structured digitisation workflows – particularly those integrating AI-driven tools, such as Intelligent Document Processing (IDP) – to automate metadata extraction, streamline validation, and enhance repository interoperability across African GLAM institutions (Ngulube 2024; Cox 2023).

South African library and archive services now face a dual imperative: to digitise and preserve historically significant collections which are under threat of degradation, and to ensure that these digital surrogates are accessible and interoperable within broader knowledge ecosystems. This requires solutions that are robust yet flexible, able to handle diverse formats and conditions while meeting international quality standards, such as ISO 19264-1:2021 for image quality, and Federal Agencies Digital Guidelines Initiative (FADGI) for capture process control (Zeutschel, 2023; ISO, 2021). The ISO 19264-1:2021 standard defines measurable criteria for image quality in the digitisation of reflective originals, including parameters such as spatial resolution, tonal response, colour accuracy, and geometric distortion (ISO, 2021). It provides a benchmark for evaluating whether scanning systems produce faithful, reproducible digital surrogates suitable for long-term preservation. In this study, ISO 19264-1:2021 serves as the technical foundation for verifying that both overhead and feed-scanning workflows meet international preservation standards comparable to those used by major cultural heritage institutions. Institutions must also contend with legacy systems, limited technical capacity, and uneven metadata practices. These barriers can be addressed through end-to-end digitisation workflows that integrate scanning, metadata extraction, validation, and export.

Previous research underscores the importance of standards-based digitisation frameworks for ensuring interoperability and preservation integrity. Studies by Gilliland (2014) highlight the evolution of archival description frameworks in the digital era, and more recent research by Gracy (2015) examines the integration of linked data and metadata standards within archival description. The Deutsche Forschungsgemeinschaft (DFG, 2022) provides practical digitisation guidelines that emphasise objective image-quality metrics. Complementary guidance highlights conformance-based quality management and the role of automation in QA/QC (FADGI, 2023; National Archives and Records Administration (NARA), (2023), while recent work explores AI-enabled approaches in heritage workflows (Schellnack-Kelly, 2024). However, literature addressing the combined implementation of ISO-aligned imaging with IDP in African archival contexts remains limited, motivating the present study.

2. Problem statement

Across African Galleries, Libraries, Archives and Museums (GLAM) institutions, digitisation projects frequently separate high-fidelity imaging from downstream metadata generation and repository integration, resulting in inconsistent image quality, fragmented metadata, and weak interoperability. This study addressed that gap by evaluating end-to-end, standards-aligned workflows that integrate ISO 19264-1:2021-compliant imaging with IDP for automated extraction, validation, and export.

While numerous digitisation initiatives have been undertaken across African libraries and archives, few have implemented truly end-to-end workflows that combine high-fidelity imaging, IDP, and standards-based metadata export. The absence of such integrated frameworks results in inconsistent image quality, fragmented metadata, and limited

interoperability between systems such as DSpace, Access to Memory (AtoM), and SharePoint. Consequently, the capacity of institutions to ensure both preservation integrity and discoverability remains constrained. Addressing this gap requires an empirical demonstration of workflows that align with international imaging standards (ISO 19264-1: 2021) and are locally implementable using AI-enabled metadata-extraction processes.

Despite growing global interest in digitisation, African archives and libraries remain under-represented in the literature on intelligent workflow automation. Existing studies rarely explore the practical integration of IDP in environments characterised by constrained resources, fragmented infrastructures, and legacy analogue holdings. This study, therefore, examined two real-world South African digitisation projects to evaluate whether IDP-enhanced workflows can reliably produce standards-compliant metadata outputs at scale. The findings have significance for both operational planning and guiding future investment in digital preservation, access, and automation infrastructure across the continent.

The purpose of this study is to evaluate the effectiveness of integrated digitisation workflows that combine archival-grade imaging systems with IDP technologies for automated metadata extraction, validation, and export. Through two applied case studies – university theses and historical newspapers – the research assessed how such workflows can enhance efficiency, accuracy, and compliance with international standards, contributing to a scalable model for standards-aligned, AI-enabled digitisation applicable to African library and archival institutions.

3. Research purpose

The purpose of this study was to investigate end-to-end digitisation workflows with intelligent document processing for university theses and historical newspapers. This study evaluated whether end-to-end digitisation workflows that combine ISO 19264-1:2021-compliant imaging with IDP can deliver measurable gains in accuracy, efficiency, and interoperability for African GLAM institutions.

4. Research objectives

The objectives of the study were as follows:

- *Accuracy and efficiency:* To quantify the effect of IDP-enabled workflows on metadata-extraction accuracy and throughput, relative to conventional practice, using operational indicators (e.g., field-level agreement with ground truth; precision/recall where applicable; pages/hour; rescans %; operator interventions per 1,000 pages).
- *Interoperability:* To assess the extent to which ISO 19264-1:2021-aligned imaging and automated metadata export support loss-free ingest and schema conformity across DSpace, SharePoint, and AtoM, using indicators such as ingest success rate, schema/field mapping completeness, hierarchy preservation, and post-ingest validation errors.

5. Research questions

The questions of the study were as follows:

RQ1. *Accuracy and efficiency:* To what extent do IDP-enabled digitisation workflows improve field-level metadata accuracy and throughput compared with conventional manual processes?
RQ2. *Interoperability:* To what extent do integrated workflows – combining ISO 19264-1:2021-1–compliant imaging with automated metadata export – enhance cross-repository interoperability (DSpace, SharePoint, AtoM) without schema loss or hierarchy degradation?

6. Scope and evaluation note

The questions are addressed through two applied South African case studies (university theses; historical newspapers). Accuracy and efficiency are evaluated against sampled records and ground truth, as well as system logs. Interoperability is assessed through ingestion and post-ingest validation in each target repository.

7. Literature review

This section presents a literature review for the study.

7.1 Accuracy and efficiency (IDP for extraction, classification, and validation)

Recent work demonstrates that IDP, which combines OCR/HTR, computer vision, and NLP, can substantially increase throughput while maintaining or improving quality in cultural heritage and records settings. Handwritten text recognition (HTR) research demonstrates measurable access gains and scaling effects for manuscript corpora, while cautioning that model design and evaluation critically shape outcomes (Nockels, Gooding & Terras, 2024). Empirical studies on table-like, pre-printed sources show that modern ML pipelines can accurately extract structured fields from noisy handwritten logbooks, evidencing concrete efficiency benefits for downstream data curation (Prieto et al., 2023). User-oriented evaluations confirm that OCR/HTR quality has a strong impact on the perceived usefulness of digitised newspapers; improvements in recognition quality directly translate into better retrieval and increased researcher satisfaction (Kettunen et al., 2022). In a government records review, supervised ML has been shown to reduce the screening burden without sacrificing legal defensibility, an efficiency pattern relevant to archival sensitivity review more broadly (Baron, Sayed & Oard, 2022). Within archives, documenting automation via "paradata" is emerging as a best practice to preserve process transparency while enabling repeatable quality control, an essential complement to accuracy claims (Davet, Hamidzadeh & Franks, 2023). Beyond individual case studies, field reviews of IDP for unstructured documents synthesise technique clusters (layout analysis, entity extraction, validation) and report accuracy gains when ML is paired with rule-based checks (Mahadevkar, Gudadhe & Chate, 2024). Finally, archivally focused AI literacy work argues that reliability in AI-assisted workflows depends on competencies for evaluation, validation, and human oversight, again reinforcing accuracy as a socio-technical outcome rather than a model-only property (Hernández & Rockembach, 2025).

7.2 Interoperability (standards, metadata, and cross-system exchange)

Image-quality and metadata standards form the foundation for interoperability across digitisation and archival pipelines. The ISO 19264-1:2021 – Photography – Archiving systems – Image quality analysis – Part 1: Reflective originals standard defines measurable parameters such as spatial resolution, tone reproduction, signal-to-noise ratio and geometric distortion, ensuring that digital surrogates accurately represent physical originals and remain verifiable across institutions (ISO, 2021.). Complementary guidance is provided by the FADGI (2023), which specifies still-image quality metrics and conformance tiers for cultural heritage digitisation. Together, these frameworks establish reproducible benchmarks for capture quality and technical validation.

The DFG Practical Guidelines on Digitisation (2022) extend these principles to project management, metadata documentation and preservation planning, emphasising institutional accountability and the long-term accessibility of digital objects (DFG, 2022.). Metadata interoperability further depends on community standards, such as the Dublin Core Collection Description Application Profile (DCMI, n.d..), which supports the consistent representation of collection-level context across repositories.

Authority control and identity alignment initiatives, such as *IdRef*, *Paprika,* and *Qualinka,* have demonstrated how structured authority data enhance discoverability and linking across systems (Le Provost & Nicolas, 2020). Within archival description, the Records in Contexts (RiC) conceptual model, developed by the International Council on Archives, introduces graph-based representations of relationships, provenance, and entities, providing a flexible, machine-actionable framework for cross-system integration (Pitti, Stockting & Clavaud, 2016).

Emerging scholarship on AI-enabled archives highlights the ethical and organisational dimensions of interoperability. Jaillant and Rees (2022) demonstrate that trust, transparency, and collaborative governance are crucial for ensuring that interconnected digital archives remain both technically and socially accountable. Collectively, these developments suggest that interoperability is not merely a technical goal but a sustained organisational practice encompassing data quality, ethical data sharing and standards-based collaboration.

7.3 Capacity, governance and sustainability (with emphasis on African contexts)

Across the African archival and library landscape, scholarship consistently highlights the structural constraints skills shortages, funding limitations, infrastructure deficits and policy gaps – that shape the sustainability of digital transformation initiatives. The edited volume "*Managing Digital Records in Africa"* synthesises empirical findings from multiple national contexts, identifying digital skills development, governance maturity, and standards adoption as decisive factors for enduring success (Ngoepe, 2022). Within this framework, open-science and digital-scholarship practices are crucial to building institutional capacity and ensuring equitable participation in the knowledge economy. More recently, Ngulube, Rabatseta, and Modiba (2024) document how South African university libraries are beginning to apply AI tools to information service delivery, highlighting both their transformative potential and the persistent infrastructural and policy challenges that accompany such adoption across African higher education settings.

Across the African archival and library landscape, researchers consistently highlight the structural constraints – skills shortages, funding limitations, infrastructure deficits and policy gaps – that shape the sustainability of digital transformation initiatives. The edited volume "*Managing Digital Records in Africa"* synthesises empirical findings from multiple national contexts, identifying digital skills development, governance maturity, and standards adoption as decisive factors for enduring success (Ngoepe, 2022). Within this framework, open-science and digital-scholarship practices are central to building institutional capacity and ensuring equitable participation in the knowledge economy. South African case studies provide granular insights into implementation realities. De la Porte (2019) identifies funding and staffing challenges that hinder the large-scale digitisation of cultural heritage materials. In contrast, Schellnack-Kelly and Modiba (2024) illustrate how AI-assisted audiovisual management can enhance lifecycle control when paired with institutional policy frameworks. More broadly, Modiba (2024) and related work on public-sector digitisation stress that sustainability requires not only technical expertise but also governance reform and cross-institutional cooperation.

Professional and normative frameworks also frame sustainability in terms of ethics and global impact. The IFLA Statement on Libraries and Artificial Intelligence emphasises the importance of equitable access, data protection, and workforce readiness as prerequisites for responsible AI integration (IFLA, 2020). At the intergovernmental level, UNESCO's Recommendation on the Ethics of Artificial Intelligence (2022) advances universal principles of transparency, accountability and inclusiveness, reinforcing the human-centred governance of digital infrastructures.

Finally, Bergstrom, Rieger and Schonfeld (2024) characterise the "second digital transformation" of scholarly communication as a shift towards shared infrastructure and collaborative stewardship, an evolution directly relevant to the archival sector's own pursuit of sustainable, interoperable, and standards-based ecosystems. Together, these perspectives converge on the need for coordinated policy frameworks, continuous professional development, and long-term infrastructure investment to secure the sustainability of intelligent digitisation across African memory institutions.

7.4 Conceptual research methodology

Building on these thematic insights into accuracy, efficiency, and interoperability within African digitisation initiatives, the following section outlines the conceptual research methodology guiding this study. This study adopted an applied qualitative research design situated within the interpretivist paradigm, aiming to explore how IDP can enhance the accuracy, efficiency, and interoperability of digitisation workflows in archival contexts. The research employed a case study methodology, which is well-suited to examining complex organisational processes and technological systems within their real-life institutional environments (Yin, 2018; Stake, 1995). By combining qualitative observation with technical evaluation, this approach enabled an in-depth understanding of both procedural and contextual factors shaping digital-transformation practices.

Case study research in archival and library science has long been employed to bridge theory and practice by examining how emerging technologies are embedded in institutional routines (Gilliland, 2014; Fidel, 1984). In this study, the case study design was complemented by design-science reasoning (Hevner, 2007), which focuses on developing and evaluating practical artefacts, such as IDP-enabled metadata-extraction components, within controlled

workflow environments. This aligns with the methodological traditions of action research and applied systems inquiry, both of which emphasise iterative refinement, stakeholder collaboration, and evaluative documentation (Checkland & Holwell, 1998; Wieringa, 2014).

The conceptual framework guiding this study, therefore, integrated evaluative design and interpretive analysis. The objective was not only to measure performance outcomes, such as processing speed and data accuracy, but also to interpret institutional readiness, governance maturity, and interoperability outcomes arising from the introduction of intelligent automation. This methodological synthesis supports evidence-based recommendations for sustainable implementation across archival and library digitisation initiatives in Africa.

8. Methodology

This study examined two digitisation projects designed and implemented by First Coast Technologies (trading as FCTEC) to evaluate the feasibility and effectiveness of IDP for extracting structured metadata and integrating it into archival workflows. Both projects were carried out at FCTEC's facilities in Cape Town and Johannesburg, utilising a combination of high-throughput feed scanners, archival-grade overhead scanners, image quality assurance tools, and a local cloud-hosted platform known as NEURO.AI, comprising a software stack adapted explicitly for Metadata as a Service (MaaS) delivery.

The digitisation pilot focused on two real-world collections: (1) bound volumes of printed university theses from the early 2000s, and (2) deteriorating newspaper holdings printed on acidic paper stock. These collections were chosen for their representativeness of widely present archival formats in African university and government repositories, as well as for their practical relevance to ongoing digitisation efforts at the hosting institutions. Document selection was guided by preservation urgency (i.e. degree of physical decay), demand for access, and institutional digitisation priorities. The thesis collection featured consistent metadata structures and large volumes of typed text in English, while the newspapers posed challenges, including poor physical condition, variable column structures, and multilingual content. This deliberate contrast enabled a comparative evaluation of the workflow's ability to handle both structured and semi-structured analogue inputs.

9. Presentation of the findings of the study

This section presents the findings of the study.

9.1 Case study 1: University theses

The first case study involved digitising and extracting metadata from a mixed-format corpus of postgraduate theses:

- **Born-digital corpus**: A set of 50 open-access postgraduate theses was harvested as PDF files from 10 South African university repositories. These were used to test field-level extraction accuracy and build baseline models for metadata prediction.
- **Printed corpus**: A further 100 printed theses were scanned using an InoTec SCAMAX 621 feed scanner at 300 dpi, 24-bit colour. This high-volume duplex scanner enabled throughput of up to 30,000 pages per day under the supervision of a trained operator. Output files included both uncompressed TIFF (archival master) and PDF/A (access surrogate).

Extracted metadata included core Dublin Core fields such as dc.title, dc.creator, dc.contributor.advisor, dc.date.issued, and dc.description.abstract. Records were reviewed via NEURO.AI's interface using confidence thresholds, and corrections were made manually when scores fell below 90%. All validated records were ingested into a self-hosted DSpace repository for evaluation and analysis.

9.2 Case study 2: Historical newspapers

The second case focused on fragile, bound volumes of the *Graaff-Reinet Advertiser*, a regional newspaper collection spanning the 19th and 20th centuries. Due to the physical condition of the materials, digitisation was conducted using a Zeutschel OS Q1 overhead scanner at 400 dpi, 24-bit colour. The scanner was calibrated to ISO 19264-1:2021 Level B quality standards. Digitisation was performed using Zeutschel's OmniScan software and verified with the integrated OS QM Tool, which performed Uniform Test Target (UTT) analysis, colour channel inspection, and batch QA. Scanned images were transferred to the NEURO.AI platform for further processing, which included optical character recognition (OCR), article segmentation, and experimental named entity recognition (NER). Outputs were tested in Microsoft SharePoint and AtoM, with structured metadata exported in METS format.

9.3 Document preparation and batch tracking

Each physical volume underwent archival document preparation, including (where relevant) staple removal, flattening, and surface cleaning. Minor conservation interventions (repairing tears) were performed where needed. Unique batch identifiers (UIDs) were assigned to each volume and linked to corresponding digital outputs. Project workflows and metadata lineage were tracked using Zoho Projects, allowing for real-time auditing and error tracking throughout the digitisation chain.

9.4 Scanning infrastructure and configuration

Two scanner types were used:

- **Feed scanner**: The InoTec SCAMAX 621, configured for 300 dpi, colour duplex scanning of unbound theses. This system included ultrasonic double-feed detection and colour de-skew correction. Image output conformed to preservation and access formats.
- **Overhead scanner**: The Zeutschel OS Q1, configured for 400 dpi, 24-bit colour scanning of bound volumes. OmniScan software controlled the cradle settings and image capture parameters. Batch scans were validated using the OS QM Tool for resolution, sharpness, contrast, and RGB integrity.

9.5 Intelligent document processing workflow

The scanned outputs were ingested into NEURO.AI, which orchestrated the IDP process through several modular stages:

- Image pre-processing: Cropping, de-skewing, binarisation, and zonal segmentation were automatically applied before OCR. Margins and gutter shadows were suppressed where required.

- OCR: Azure Cognitive Services was used for OCR, with language models set to English and Afrikaans. Layout-aware OCR systems (e.g. Transkribus-compatible models) were not deployed, as handwritten elements were negligible.
- Article segmentation: For newspapers, OCR text blocks were segmented into logical units (headline, date, body text).
- NER (experimental): Named entity recognition models were tested on newspaper text blocks to identify people and place names. Due to low confidence levels, these outputs were not retained as formal metadata.

9.6 Classification architecture

Although automated classification into document types (e.g., thesis, article, advertisement) is supported by the workflow architecture, it was not required for these two case studies. Document class was predefined by filenames and source metadata. However, the system architecture retained classification capabilities using:

1. Layout heuristics: Page structure indicators such as headers, tables of contents, or length.
2. Filename and metadata matching: Inference based on pre-existing folder structures and labels.

Confidence scores for classification were recorded for reference but not acted upon in these projects.

10. Field-level data extraction

The following fields were extracted:

- For theses:
    - Title (dc.title)
    - Author (dc.creator)
    - Advisor (dc.contributor.advisor)
    - Institution
    - Degree
    - Date (dc.date.issued)
    - Abstract (dc.description.abstract)
    - URI (dc.identifier.uri)
- For newspapers:
    - Article text blocks
    - Headline and issue date (where identifiable)
    - Body text
    - Named entities (not formally captured due to low confidence)
- Structural metadata: Logical page order and preliminary article boundaries were retained in METS XML outputs.

11. Validation procedures

Confidence scores were automatically assigned per field by the IDP engine. Two validation strategies were used:

- ETDs: Metadata fields scoring below 90% confidence were manually reviewed using the NEURO.AI interface. Review and correction took between 20 and 180 seconds per record.
- Newspapers: Lower thresholds were tolerated due to the exploratory nature of the extraction and absence of ground-truth data. Outputs were not manually corrected but retained for future refinement.

12. Metadata normalisation and export

Final metadata conformed to recognised standards:

1. Descriptive metadata: Dublin Core (DCMI) fields were used for repository ingestion via DSpace. Mapping templates in NEURO.AI ensured consistent alignment.
2. Preservation metadata: Where applicable, PREMIS metadata documented the digitisation process, agents, and file format characteristics (TIFF, PDF/A).
3. Structural metadata: METS served as the container.

All processed outputs were exported to institutional systems hosted in sandbox environments:

- ETDs: Batch-ingested into a local instance of DSpace.
- Newspapers: Exported to SharePoint and test-ingested into AtoM. The latter encountered import challenges, particularly with hierarchical series and subseries structures using standard CSV ingest templates.

Folder-naming conventions (e.g., institutioncode_collectionID_itemID) and manifest files ensured system interoperability. NEURO.AI generated ingestion-ready packages, including metadata, image files, and audit logs.

13. Workflow control and monitoring

All processes were tracked via Zoho Projects, providing:

- workflow transparency
- operator accountability
- batch-level reporting
- error and correction tracking
- end-to-end audit trail from scan to repository ingest.

This ensured adherence to internal QA standards and provided a complete record of human and machine intervention across both case studies.

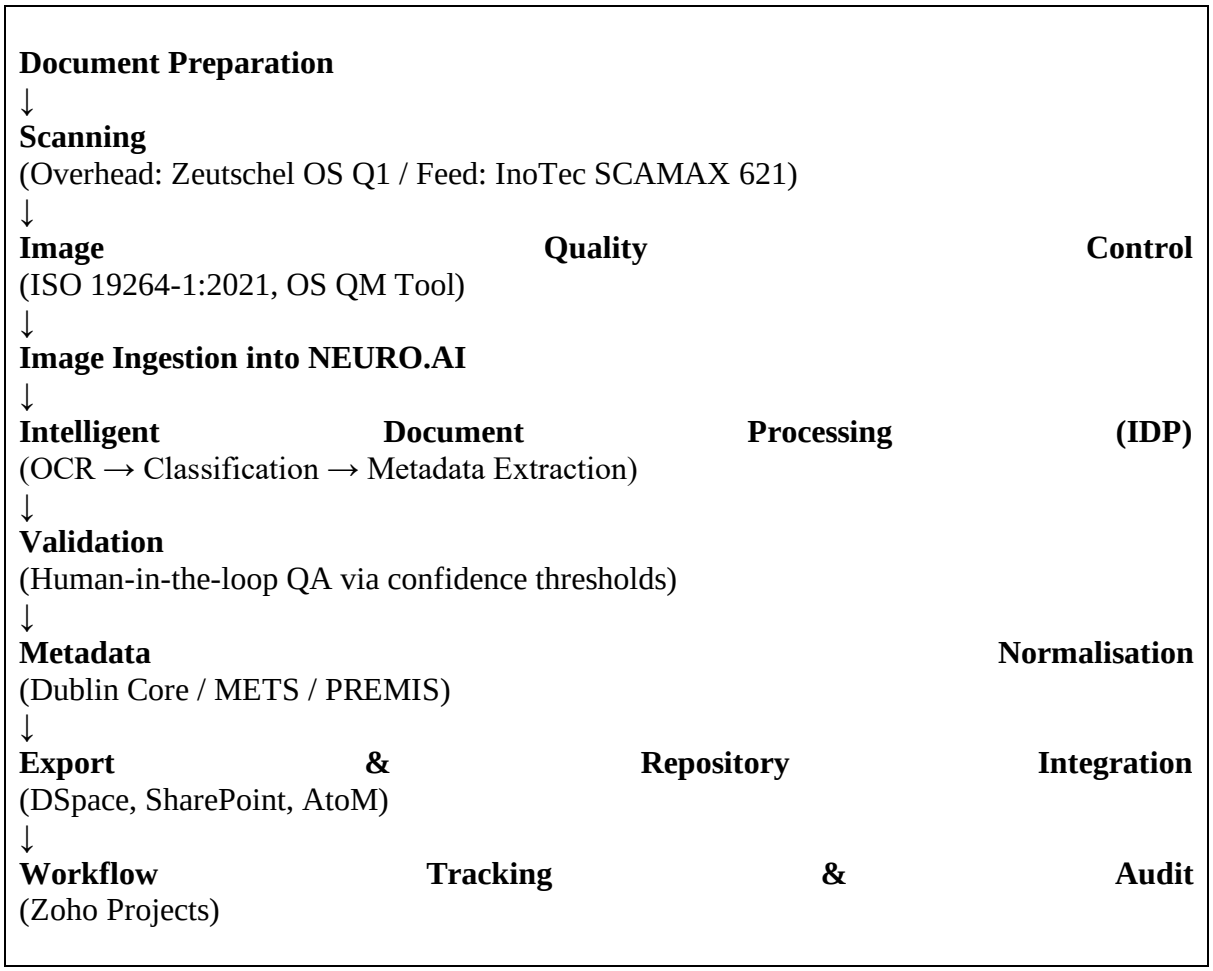

Figure 1. Overview of the end-to-end digitisation and IDP workflow applied in both case studies, showing the integration of scanning, image quality control, metadata extraction, validation, and repository export.

14. Interpretation and analysis of the results

The implementation of structured, end-to-end digitisation workflows across two contrasting archival material types – printed academic theses and historical newspapers – yielded insights into throughput scalability, metadata-extraction reliability, and the functional utility of IDP in archival contexts. Both projects were evaluated not only for their technical feasibility but also for their adherence to archival standards, integration with existing institutional platforms, and potential to enhance the long-term discoverability and reuse of digital surrogates. Results are presented per case study and structured around throughput performance, extraction accuracy, metadata integration, and system interoperability.

14.1 University theses and dissertations

A corpus of 100 printed theses, comprising approximately 150,000 pages, was digitised over five working days at FCTEC's Johannesburg bureau. The InoTec SCAMAX 621 feed scanner, operating at 300 dpi and 24-bit colour, consistently achieved an average daily throughput of approximately 30,000 pages, aligning with manufacturer specifications and prior empirical benchmarks (InoTec, 2023). Automated feed correction and ultrasonic misfeed detection reduced manual intervention, with fewer than 0.5% of pages requiring rescanning due to skew, truncation, or double feeds.

Post-scan ingestion into the NEURO.AI software stack triggered automated IDP pipelines for OCR, layout parsing, and field-level metadata extraction. The target schema consisted of nine core fields aligned with Dublin Core and South African ETD repository schemas: dc.title, dc.creator, dc.contributor, dc.subject, dc.date, dc.identifier, dc.type, dc.language, and dc.rights. These fields reflect common practice across OpenUCT, SUNScholar, and WIReDSpace repositories. Confidence levels reported by the system exceeded 90 % across dc.title, dc.creator, and dc.date, with slightly lower but acceptable results for dc.subject and dc.identifier.

Independent manual validation of a 20% random sample confirmed strong correspondence between the extracted and ground-truth metadata values. Although system confidence metrics serve as useful heuristics, they are not direct substitutes for precision-recall calculations, which require benchmarked datasets and confusion matrix analysis. Nonetheless, observed mismatches were rare and largely attributable to OCR artefacts on title pages or footers.

The full-text OCR layer achieved character-level recognition accuracy of approximately 98%, as determined by a Levenshtein-distance comparison between the OCR output and known abstracts for selected theses. Even where pages contained marginalia, embedded diagrams, or watermarks, the pipeline maintained high fidelity. Outputs were structured into PDF/A and METS-encoded packages and successfully batch-ingested into a sandbox DSpace 7.6 instance, with XML crosswalks confirming schema compliance. No ingest errors or compound-object failures occurred, demonstrating seamless interoperability between the IDP pipeline and repository backend.

14.2 Historical newspapers

A total of 785 pages from the *Graaff-Reinet Advertiser* were digitised using a Zeutschel OS Q1 planetary scanner at FCTEC's Cape Town facility. The collection included both bound volumes and loose-leaf editions dating from 1869 to 1942, with fragile paper and fading inks typical of iron-gall and wood-pulp degradation. The Zeutschel's motorised cradle and precision-guided glass plate enabled curvature-free imaging without Newton-ring artefacts or gutter shadows. Image quality was validated via Zeutschel's OS QM Tool and Uniform Test Targets (UTT-19264), benchmarked against ISO 19264-1:2021 Level B parameters. All scans met or exceeded thresholds for resolution, sharpness, tonal response, and geometric fidelity.

A subset of 100 pages underwent advanced NEURO.AI processing for automated article segmentation and NER. Article segmentation achieved a success rate of approximately 90% when benchmarked against a manually tagged control set. Boundary errors occurred primarily on multi-column layouts or pages with irregular typography. NER targeted personal, place, and

institutional names, achieving high precision for unambiguous entities such as "Sir Andries Stockenström," but reduced accuracy in degraded or bilingual text. Because no gold-standard-labelled dataset exists for Afrikaans-English newspapers of this era, NER results were retained for experimental analysis only.

Processed outputs were exported to two repository environments: a Microsoft SharePoint archive and a local instance of AtoM (version 2.7). SharePoint ingestion accepted full file sets, including METS metadata, without schema conflicts. Basic search and filter functions over OCR-derived metadata proved effective, though limited by SharePoint's lack of hierarchical archival description. AtoM imports succeeded technically but lost multi-level hierarchy across fonds, series, file, and item layers due to CSV-to-EAD template constraints – a known limitation of AtoM's bulk importer rather than of the NEURO.AI export format.

14.3 Cross-case observations

Across both case studies, embedding IDP within standards-compliant digitisation workflows delivered measurable efficiencies in throughput, metadata capture, and repository interoperability. While automation did not eliminate the need for human validation, particularly with irregular or degraded source materials, it substantially reduced manual input in the scanning and metadata-entry stages. The consistent success in exporting to downstream systems despite differing architectures demonstrates the modular resilience of well-structured pipelines. Collectively, the results indicated that hybrid digitisation with IDP integration provides a practical foundation for scalable MaaS implementations in African GLAM contexts.

15. Discussion and recommendations

The discussion evaluates the two research questions by examining, firstly, how effectively IDP-enabled workflows improve metadata-extraction accuracy and throughput, and, secondly, to what extent these workflows enhance interoperability across repositories. The results affirmed the technical viability and strategic utility of modular, standards-based digitisation within South African archival and academic institutions.

15.1 Strategic and practical implications

A key finding across both cases is the demonstrable viability of hybrid workflows that unite archival-grade imaging with automated extraction and export. These do not simply streamline digitisation; they embed machine-readability and metadata structure at the point of creation – a departure from the manual post-processing that has long dominated in regional archives. The consistent 90%+ confidence levels for core Dublin Core fields show that machine-assisted indexing can supplement, and in structured collections even replace, routine cataloguing.

Such gains parallel findings in international heritage projects where intelligent automation enhances throughput and consistency (Ding et al., 2021; Muehlberger et al., 2019). In African settings, this is particularly valuable given chronic staff shortages and cataloguing backlogs (Ngoepe, 2022; IFLA, 2020). By automating first-stage enrichment, professionals can focus on high-value review and contextual validation, following the "human-in-the-loop" paradigm emphasised in recent guidance on algorithmic literacy and human-centred AI (Ridley & Pawlick-Potts, 2021; UNESCO, 2021). The successful DSpace integration further evidences practical interoperability, leveraging non-proprietary schemas such as Dublin Core (and, for

archival systems, EAD/ISAD(G) alignment). This aligns with UNESCO's (2021) emphasis on human-centric, rights-based and inclusive approaches to digital infrastructure development, particularly in contexts of the Global South.

15.2 Standards compliance and interoperability

Both workflows complied with ISO 19264-1:2021 Level B, confirming that preservation-grade imaging is achievable in African contexts using commercial equipment and skilled operation. Validation through Zeutschel's QM Tool and UTT-19264 targets verified resolution, tonal accuracy, and geometric fidelity, aligning with best practices at the Library of Congress and the National Archives of the Netherlands. This principle is echoed locally: FCTEC (2023) emphasises that digitisation in African GLAM settings should align with widely adopted standards to support longevity, accuracy, and collaboration, reinforcing the case for ISO 19264-1:2021-aligned capture and formal metadata schemas.

Metadata generated by IDP was successfully integrated into open-source repositories. DSpace ingestion proceeded without loss, whereas AtoM imports revealed known structural limitations. Adherence to open metadata standards such as Dublin Core and EAD supports long-term sustainability by avoiding vendor lock-in and enabling migration, harvesting, and cross-repository syndication. Crucially, standards-based interoperability facilitates regional collaboration, allowing shared infrastructure and automation services among capacity-limited institutions.

15.3 Confidence and human oversight

While system-generated confidence levels above 90% are promising, they remain algorithmic estimates rather than empirically validated metrics, such as precision, recall, or F1 score. As Kettunen et al. (2022) demonstrate, OCR-generated confidence values provide useful heuristics but do not reliably indicate true metadata accuracy, particularly in multilingual or noisy historical sources, where layout and language variability still produce systematic recognition errors. Manual validation of representative samples provided a secondary check, but future research should develop gold-standard corpora to enable objective benchmarking.

Material variability also influences accuracy: structured theses yielded more reliable results than complex newspaper layouts, echoing Ehrmann, Bunout and Clavert (2022), who note that heterogeneous typography, page segmentation inconsistencies, and historical print degradation continue to limit OCR reliability and downstream data extraction in digitised newspaper collections. Persistent errors in article segmentation and NER confirm that human expertise remains essential. This aligns with international guidance that emphasises human oversight and professional accountability in cultural heritage AI. UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021) frames AI as supporting, rather than supplanting, human judgement, and European policy briefings on AI in cultural heritage similarly advocate for a human-centric approach to deployment. For African institutions, limited staffing underscores the need for targeted training, validation protocols, and in-country partnerships to enhance data governance and reduce dependence on opaque global platforms.

15.4 Broader implications and future directions

The study supports the view that contemporary digitisation strategies are shifting from preservation-only aims towards the creation of machine-actionable, interoperable data assets – what recent scholarship on digitised newspapers calls the transformation of sources into "complex data objects" (Ehrmann et al., 2022; Michael, 2024), with clear implications for access and interoperability in African memory institutions (De la Porte, 2019). Demonstrated capacity to produce standards-compliant, repository-ready outputs underscores the feasibility of implementing MaaS frameworks even within constrained environments. The deployment of NEURO.AI, a locally developed and hosted platform, exemplifies how regional innovation can ensure both technical autonomy and responsible data stewardship.

Nevertheless, structural challenges persist, including inadequate metadata frameworks, policy gaps, and limited technical capacity, which continue to hinder large-scale sustainability (Mutula, 2010). Future research should (1) create benchmark datasets for African archival materials, (2) evaluate the generalisability of IDP pipelines for handwritten and multilingual sources, and (3) assess the return on investment (ROI) of MaaS frameworks in institutional and national archives.

By combining automation, open standards, and professional oversight, scalable digitisation becomes both technically feasible and institutionally sustainable – a pathway towards resilient digital transformation across the African GLAM sector.

16. Conclusion

This study demonstrated the practical viability and strategic importance of end-to-end digitisation workflows that integrate archival-grade imaging with IDP. Using two South African case studies (academic theses and historical newspapers), it was demonstrated that modular, standards-based systems can deliver both high-quality preservation outputs and enriched, interoperable metadata suitable for repository integration.

The application of international imaging benchmarks, such as ISO 19264-1:2021, and open metadata schemas, like Dublin Core, confirmed that preservation-grade digitisation is attainable with commercially available hardware and software, provided the workflow is underpinned by calibration, quality assurance, and human-in-the-loop validation. This finding aligns with earlier recommendations by the DFG (2022) and the FADGI (2023), which emphasise measurable image-quality standards as central to digital preservation integrity.

At a strategic level, the integration of IDP reframes digitisation as an ongoing process of data enrichment rather than a terminal conversion act. Similar perspectives have been advanced by Colavizza et al. (2021) and Hernández and Rockembach (2025), who argue that AI-assisted archival systems enhance metadata quality and institutional accountability. When aligned with institutional repositories, such as DSpace, or shared storage solutions, like SharePoint, intelligent workflows enhance discoverability and interoperability across systems.

Looking forward, three priorities emerge. Firstly, the empirical evaluation of IDP outputs using precision-recall metrics would enable more rigorous benchmarking of accuracy across African record types. Secondly, developing and training models for handwritten, multilingual, and region-specific materials would help address current data bias gaps. Thirdly, expanding MaaS

frameworks could offer cost-effective, cloud-enabled pathways for metadata enrichment and interoperability in shared repositories.

Ultimately, the combination of intelligent automation, open standards, and expert human validation provides a credible pathway for sustainable digitisation in Africa's galleries, libraries, archives, and museums. This study presents one of the first empirically documented demonstrations of ISO 19264-1:2021-aligned intelligent digitisation workflows within an African archival context, offering both methodological insights and practical guidance for GLAM institutions undertaking digital transformation. The successful DSpace integration further evidences practical interoperability, leveraging non-proprietary schemas such as Dublin Core and ISAD(G). This aligns with the principles of the IFLA/UNESCO Manifesto for Digital Libraries (2011), which promotes sustainable and interoperable digitisation, and with UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021), emphasising inclusive, human-centred digital infrastructures.

Declaration

- The manuscript has not been previously published and is not under consideration for publication with any other journal or copyrighted publishing platform of any kind.
- Permission was granted for collection and publication of the presented identifiable data from target organisation (Permission letter will be provided with manuscript).
- Unlawful statements that infringe any existing copyrights is avoided in the manuscript.
- Author have sought permission for copyright from 3rd parties by obtaining the necessary permission from the copyright holder/s to reproduce their materials in different media in the article, such as tables, diagrams and photographs owned by them.
- Manuscript and study meet all the ethical requirements of the journal and that of my institution or company, as well as legal requirements of the study country.
- No AI tool/Large Language Model (LLM) was used in the creation of manuscript. Any usage of AI tool/LLM in the development or generation of manuscript in part or in full is stated clearly in the Methodology and Acknowledgements sections of the manuscript.
- There is not any potential conflict of interest for the research.
- Author is familiar with the content of this manuscript and gave consent to co-publish.
- Authors contributed to the writing of the article manuscript.
- Author takes responsibility to keep participants information confidential as required by legislations including Protection of personal information Act.
- Author gives consent to the Journal of South African Society of Archivist to publish the manuscript.

Acknowledgements

The author declares that no external funding, institutional support or third-party contributions were received for this study.

References

Balado, A.B., de Boer, V. & Schreiber, G. 2015. Linking historical ship records to a newspaper archive. In Aiello, L. M. & McFarland, D. (eds.), *Social Informatics: SocInfo 2014 Workshops,* Barcelona, 11 Nov 2014, Revised Selected Papers. Cham: Springer, pp. 254–263. https://doi.org/10.1007/978-3-319-15168-7_28

Balasubramanian, S. & Tamilselvan, N. 2023. Exploring the potential of artificial intelligence in library services: A systematic review. *International Journal of Library & Information Science,* 12(1).

Baron, J.R., Sayed, M.F. & 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): 1–19. https://doi.org/10.1145/3481045

Bergstrom, T., Rieger, O.Y. & Schonfeld, R.C. 2024. *The second digital transformation of scholarly publishing: Strategic context and shared infrastructure.* New York: Ithaka S+R. https://doi.org/10.18665/sr.320210

Colavizza, G., Blanke, T., Jeurgens, C. & Noordegraaf, J. 2021. Archives and AI: An overview of current debates and future perspectives. *ACM Journal on Computing and Cultural Heritage,* 15(1): 4. https://doi.org/10.1145/3479010

Cox, A.M. 2023. How artificial intelligence might change academic library work: Applying the competencies literature and the theory of the professions. *Journal of the Association for Information Science and Technology,* 74(3): 367–380. https://doi.org/10.1002/asi.24635

Davet, J., Hamidzadeh, B. & Franks, P. 2023. Archivist in the machine: Paradata for AI-based automation in the archives. *Archival Science.* Advance online publication. https://doi.org/10.1007/s10502-023-09408-8

De la Porte, B. 2019. Challenges in digitisation of cultural heritage material in memory institutions in the Western Cape, South Africa. *South African Journal of Libraries and Information Science,* 85(1): 1–12. Available at:https://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S1560-683X2019000100030 (accessed 12 April 2025)

Deutsche Forschungsgemeinschaft (DFG). 2022. *Practical guidelines on digitisation (updated version).* Bonn: DFG. Available at: https://www.dfg.de/resource/blob/176110/76abec10bdc30b41f18695145003d6db/12-151-v1216-en-data.pdf (accessed 12 April 2025)

Ding, W., Zhang, J., Lin, X. & Chen, Y. 2021. AI-driven cultural heritage digitisation: Improving efficiency and quality through intelligent automation. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP),* pp. 2905–2909. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICIP42928.2021.9506202

Dublin Core Metadata Initiative (DCMI). n.d. *CDP Dublin Core metadata best practices Version 2.1.* Available at: https://sustainableheritagenetwork.org/digital-heritage/cdp-dublin-core-metadata-best-practices-version-21 (accessed 12 April 2025)

Dublin Core Metadata Initiative (DCMI). n.d. *Dublin Core™ collection description application profile: Data Model.* Available at: https://www.dublincore.org/specifications/dublin-core/collection-description/collection-model/ (accessed 12 April 2025)

Ehrmann, M., Bunout, E. & Clavert, F. 2022. Digitised historical newspapers: A changing research landscape. In Brügger, N. & Laursen, D. (eds.), *Digital histories?* Berlin: De Gruyter, pp. 1–19. https://doi.org/10.1515/9783110729214-001

Fidel, R. 1984. The case study method: A case study. *Library & Information Science Research*, 6(3): 273–288. Available at: https://faculty.washington.edu/fidelr/RayaPubs/TheCaseStudyMethod.pdf (accessed 12 April 2025)

First Coast Technologies (Pty) Ltd. 2023. *The importance of standards in digitising Africa's cultural heritage.* Midrand: FCTEC. Available at: https://fctec.co.za/blog/the-importance-of-standards-in-digitising-africas-cultural-heritage/ (accessed 12 April 2025)

Federal Agencies Digital Guidelines Initiative (FADGI). 2023. *Technical guidelines for digitizing cultural heritage materials* (3rd ed.). Washington, D.C.: U.S. Library of Congress. Available at: https://www.digitizationguidelines.gov/guidelines/FADGI%20Technical%20Guidelines%20for%20Digitizing%20Cultural%20Heritage%20Materials_3rd%20Edition_05092023.pdf (accessed 12 April 2025)

Gilliland, A.J. 2014. Conceptualising 21st-century archives. In Eastwood, T. & MacNeil, H. (eds.), *Currents of archival thinking* (2nd ed.). London: Facet, pp. 607–634.

Gracy, K.F. 2015. Archival description and linked data: A preliminary study of opportunities and implementation challenges. *Archival Science*, 15(3): 239–294. https://doi.org/10.1007/s10502-014-9216-2

Hernández, R.A. & Rockembach, M. 2025. Building trustworthy AI solutions: Integrating artificial-intelligence literacy into records management and archival systems. *AI and Society*, 40(6): 4265–4282. https://doi.org/10.1007/s00146-025-02194-0

Hevner, A.R. 2007. A three-cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2): 87–92. Available at: https://aisel.aisnet.org/sjis/vol19/iss2/4 (accessed 12 April 2025)

InoTec GmbH Organisation und Automatisierungstechnik. 2023. *SCAMAX 621 Series — High-Performance Production Scanners.* Wetzlar: InoTec GmbH. Available at: https://www.inotec.eu/en/products/scamax-621-series (accessed 12 April 2025)

InoTec GmbH Organisationssysteme. n.d. *SCAMAX® Scan+ software – Advanced scanning solution.* Available at: https://www.inotec.eu/scanclients/?lang=en (accessed 12 April 2025)

International Council on Archives (ICA). 2025. *AI and archival practice: A video series exploring the application of AI to the management of archives.* Available at: https://www.youtube.com/watch?v=TwKPxHUD0nM (accessed 12 April 2025)

International Federation of Library Associations and Institutions (IFLA). 2020. *IFLA Statement on Libraries and Artificial Intelligence.* The Hague: IFLA. Available at: https://repository.ifla.org/handle/123456789/1646 (accessed 12 April 2025)

International Federation of Library Associations and Institutions (IFLA)/United Nations Educational, Scientific and Cultural Organization (UNESCO). 2011. *IFLA/UNESCO Manifesto for Digital Libraries.* The Hague: IFLA. Accessed at: https://repository.ifla.org/items/dc610163-cbec-4c02-a3f4-ff42585f0e12 (accessed 12 April 2025)

International Organization for Standardization (ISO). 2021. *ISO 19264-1:2021 — Photography — Archiving Systems — Imaging Systems Quality Analysis — Part 1: Reflective Originals.* Geneva: ISO. https://www.iso.org/standard/79172.html (accessed 12 April 2025)

Jaillant, L. & Rees, A. 2022. Applying AI to digital archives: Trust, collaboration and shared professional ethics. *Digital Scholarship in the Humanities*, 38(2): 621–638. https://doi.org/10.1093/llc/fqac073

Kettunen, K., Keskustalo, H., Kumpulainen, S., Pääkkönen, T. & Rautiainen, J. 2022. OCR quality affects perceived usefulness of historical newspaper clippings — A user study. *arXiv* preprint, arXiv:2203.03557. Available at: https://arxiv.org/abs/2203.03557 (accessed 12 April 2025)

Le Provost, A. & Nicolas, Y. 2020. IdRef, Paprika and Qualinka: A toolbox for authority data quality and interoperability. *ABI Technik*, 40(2): 158–168. https://doi.org/10.1515/abitech-2020-2006

Lee, B.C.G. 2023. The 'Collections as ML Data' checklist for machine learning and cultural heritage. *Journal of the Association for Information Science and Technology*, early view. https://doi.org/10.1002/asi.24765

Mahadevkar, S.V., Gudadhe, M. & Chate, P. 2024. Exploring AI-driven approaches for unstructured document processing: A review. *Journal of Big Data*, 11: 78. https://doi.org/10.1186/s40537-024-00948-z

Michael, H. 2024. Digital press archives for media and communication history: Potentials and challenges. Preprint. University of Bamberg. Available at: https://fis.uni-bamberg.de/bitstreams/06f01a37-13e0-4fb8-ba33-a20714590fa1/download (accessed 12 April 2025)

Modiba, M. 2024. Adoption of artificial intelligence to enhance records management practices at Gauteng Department of Education in South Africa. *Collection and Curation*, 44(1): 9–17. https://doi.org/10.1108/CC-12-2023-0044

Muehlberger, G., Seaward, L., Springmann, U., Terras, M. & Kahle, P. 2019. Transforming scholarship in the archives through handwritten text recognition – Transkribus as a case study. *Journal of Documentation*, 75(5): 954–976. https://doi.org/10.1108/JD-07-2018-0114

Mutula, S.M. 2010. Digital economies in Africa: Evidence from the digital divide. *Information Development*, 26(2): 144–153. https://doi.org/10.1177/0266666910366409

National Archives and Records Administration (NARA). 2023. Technical Guidelines for Digitizing Archival Materials. Washington, DC. Available at: https://www.archives.gov/preservation/technical/guidelines (Accessed 10 December 2025)

Ngoepe, M. (ed.). 2022. *Managing digital records in Africa*. London: Routledge. https://doi.org/10.4324/9781003203155

Ngulube, P. 2024. Artificial intelligence and the future of archives and records management in Africa. *Journal of the South African Society of Archivists*, 57: 1–15.

Ngulube, P., Rabatseta, P.C. & Modiba, M. 2024. Utilisation of artificial intelligence for the provision of information services at the University of Limpopo libraries. *South African Journal of Libraries and Information Science*, 90(2). https://doi.org/10.7553/90-2-2394

Nockels, J., Gooding, P. & Terras, M. 2024. The implications of handwritten text recognition for accessing the past at scale. *Journal of Documentation*, 80(7): 148–167. https://doi.org/10.1108/JD-09-2023-0183

Pitti, D., Stockting, B. & Clavaud, F. 2016. An introduction to "Records in Contexts": An archival description draft standard. *Comma*, 2016(1–2): 173–188.

Prieto, J.R., Cano, E., Nieto, J., Santos, A. & Rodríguez, R. 2024. Information extraction in handwritten historical logbooks. *Pattern Recognition Letters*, 172: 128–136. https://doi.org/10.1016/j.patrec.2023.06.008

Ridley, M. & Pawlick-Potts, D. 2021. Algorithmic literacy and the role for libraries. *Information Technology and Libraries,* 40(2): 2. https://doi.org/10.6017/ital.v40i2.12963

Schellnack-Kelly, I. 2024. Leveraging artificial intelligence for managing audiovisual lifecycle and access in heritage institutions. *Journal of Information Science* (advance online publication). https://doi.org/10.1177/01655515241263327

Schellnack-Kelly, I. & Modiba, M. 2024. Developing smart archives in Society 5.0: Leveraging artificial intelligence for managing audiovisual archives in Africa. *Information Development.* https://doi.org/10.1177/02666669241286224

Stake, R.E. 1995. *The art of case study research*. Thousand Oaks, CA: Sage. Available at: https://uk.sagepub.com/en-gb/eur/the-art-of-case-study-research/book4954 (accessed 12 April 2025)

UNESCO. 2021. *Recommendation on the ethics of artificial intelligence*. Paris: United Nations Educational, Scientific and Cultural Organization. Available at: https://digitallibrary.un.org/record/4062376?ln=en&v=pdf (accessed 12 April 2025)

Wieringa, R.J. 2014. *Design science methodology for information systems and software engineering*. London: Springer. https://doi.org/10.1007/978-3-662-43839-8

Yin, R.K. 2018. *Case study research and applications: Design and methods*. 6th ed. Thousand Oaks, CA: Sage. Available at: https://uk.sagepub.com/en-gb/eur/the-art-of-case-study-research/book4954 (accessed 12 April 2025)

Zeutschel GmbH. n.d. *A consistently high image quality can only be achieved with standards.* Available at: https://www.zeutschel.de/en/a-consistently-high-image-quality-can-only-be-achieved-with-standards/(accessed 12 April 2025)

Zeutschel GmbH. n.d. *OmniScan software – Professional scanning software.* Available at: https://www.zeutschel.de/en/software/ (accessed 12 April 2025)

Zeutschel GmbH. 2023. *OS QM Tool — Image Quality Management Software for FADGI and ISO 19264.* Tübingen: Zeutschel GmbH. Available at: https://www.zeutschel.com/en/products/os-qm-tool/(accessed 12 April 2025)

Zeutschel GmbH. 2024. *OS Q1 — High-resolution planetary scanner for books and documents.* Tübingen: Zeutschel GmbH. Available at: https://www.zeutschel.com/en/products/os-q1/(accessed 12 April 2025)

Zhang, S., Hou, J., Peng, S., Li, Z., Hu, Q. & Wang, P. 2023. ArcGPT: A large language model tailored for real-world archival applications. *arXiv* preprint, arXiv:2307.14852. Available at: https://arxiv.org/abs/2307.14852 (accessed 12 April 2025)