Microsoft

RIFUMO
EMPOWERMENT HOLDINGS (PTY) Ltd

Digitization

Table of Contents

1 Executive Summary

1.1 Introduction

This proposal outlines an end-to-end digitization and intelligent document management solution that captures paper and born-digital content, applies OCR and AI-based extraction with Azure AI Document Intelligence, persists content and metadata into Microsoft Fabric (OneLake) for analytics, and exposes secure, enterprise search through Azure AI Search. Long-term document management, records controls, and compliant archiving are delivered through OpenText, with seamless integration into Microsoft 365 business workflows.

1.2 Business Drivers & Outcomes

Reduce manual document handling and data capture effort by 40–60% through automated classification and extraction.

Accelerate turnaround times (intake-to-decision) with straight-through processing and human-in-the-loop validation for exceptions.

Establish a governed document and data estate: authoritative records in OpenText, curated data in Fabric, and enterprise search across both.

Enable operational transparency via Power BI dashboards for scanning throughput, SLA compliance, quality, and workload balancing.

Ensure compliance (retention, legal hold, privacy) with OpenText records management and Microsoft Purview metadata and lineage.

1.3 In Scope

Document capture from scanners/MFDs, bulk import, secure email, and API-based ingestion.

AI-powered OCR and data extraction using Azure AI Document Intelligence (prebuilt and custom models).

Content storage in OpenText with metadata synchronization and records management.

Data landing and curation in Microsoft Fabric OneLake (Delta) for analytics and reporting.

Enterprise search using Azure AI Search with hybrid keyword + vector semantics and security trimming.

Power BI operational dashboards and quality monitoring.

Security, governance, and DevSecOps foundations across environments (Dev/Test/Prod).

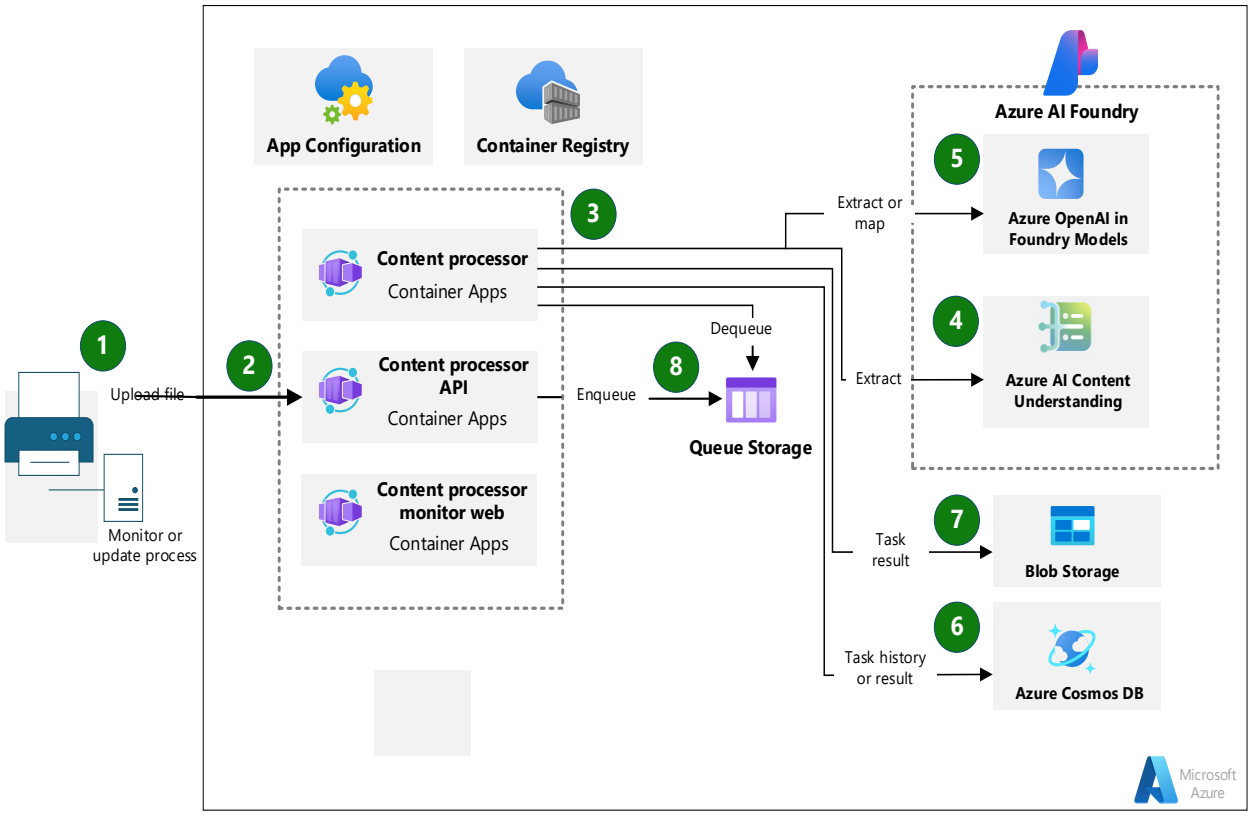
1.4 Out of Scope (initial phase)

Robotic process automation of downstream LOB transactions (can be addressed in a subsequent workstream).

Legacy document backfile conversion beyond the prioritized backlog (option available).

Process re-engineering outside of document-centric steps.

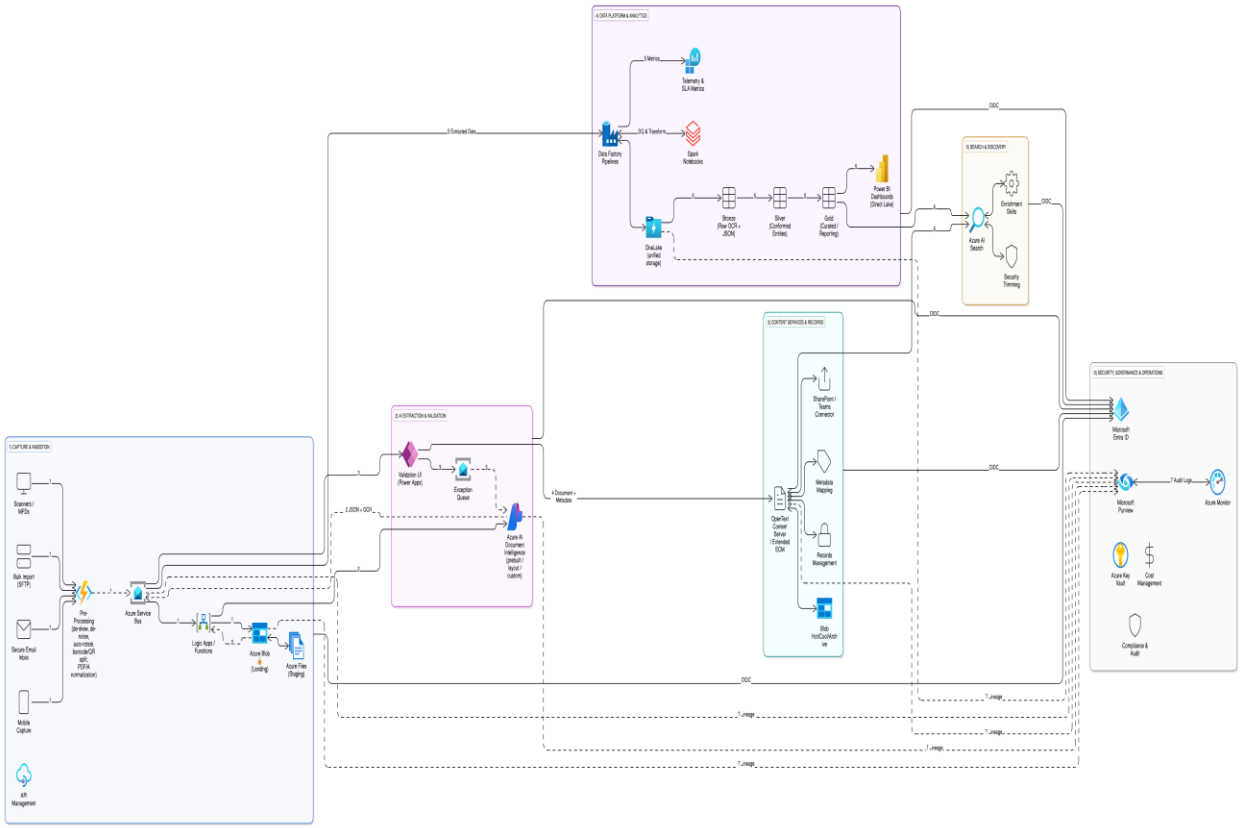
2 Solution Overview



2.1 End-to-End Value Chain

- Capture & Ingestion: Scan/import documents, apply basic QA (page order, de-skew, split/merge).
- Classification & OCR: Use AI models to classify document types and extract text, tables, and key fields.
- Validation: Human-in-the-loop verification of low-confidence fields; feedback loop for model retraining.
- Persistence: Store authoritative documents and metadata in OpenText; land extracted data in Fabric OneLake.
- Index & Search: Build Azure AI Search indexes with semantic ranking and vector embeddings for secure enterprise search.
- Consume & Govern: Business users work in OpenText and Microsoft 365; analytics in Power BI; governance via Purview and OpenText Records.

2.2 Target Architecture



- Capture Layer: High-speed scanners/MFDs, email ingestion, SFTP/API gateways; standardized profiles (PDF/A-2, 300 DPI).
- AI Extraction: Azure AI Document Intelligence (layout, prebuilt invoices/IDs/tax forms, custom neural models).
- Content Services: OpenText Content Server / Extended ECM for Microsoft 365 as the system of record (versioning, metadata, records, legal hold).
- Data Platform: Microsoft Fabric OneLake with Lakehouse/Warehouse for curated, analytics-ready datasets (Delta).
- Search & Discovery: Azure AI Search with hybrid BM25 + vector (embeddings), semantic ranker, OCR text enrichment, and filters.
- Integration & Apps: Power Automate/Logic Apps for orchestration; Power Apps for validation UI; connectors to ERP/CRM/case systems.
- Security: Microsoft Entra ID (RBAC), Key Vault (secrets, CMK), private endpoints, network isolation, and audit logs.

2.3 Process Flows (High Level)

Scan/ingest → pre-processing (de-skew, noise removal, barcode split) → blob landing.

Trigger Azure Functions/Logic Apps → call Document Intelligence → capture JSON results (fields, tables, signatures).

Route to validation app if confidence < threshold; otherwise straight-through to OpenText with metadata.

Persist extracted entities into Fabric (Bronze/Silver/Gold) → Power BI dashboards → operational insights.

Create/refresh Azure AI Search indexes referencing OpenText content locations; security trimming via application roles.

2.4 Non-Functional Requirements

- Scalability: ≥ 5 million pages/year; elastic processing via Azure Functions/AKS as needed.
- Performance: < 4 hours ingestion-to-availability for standard batches; < 5 seconds median search latency.
- Reliability: 99.9% platform availability; retry and dead-lettering for failed batches.
- Security & Compliance: POPIA/GDPR alignment; encryption at rest and in transit; data residency per policy.
- Observability: End-to-end telemetry, tracing, and cost/performance dashboards.

3 Approach & Methodology

3.1 Delivery Phases

Phase 0 – Mobilize & Foundations: Project governance, environments, connectivity, security baseline, and templates.

Phase 1 – Discover & Design: Document taxonomy, metadata model, retention schedule, sample sets, and UX prototypes.

Phase 2 – Build Platform: Deploy Azure services, OpenText integration, Fabric lakehouse, and Azure AI Search indexes.

Phase 3 – Pilot Use Cases: Configure 3–5 high-value document types; implement validation UI; measure accuracy and throughput.

Phase 4 – Scale & Industrialize: Expand to additional document classes; optimize models; automate retraining; harden ops.

Phase 5 – Operate & Optimize: Run, monitor, and continuously improve; cost governance and benefits realization.

3.2 Ways of Working

Agile delivery (2-week sprints) with product backlog, demos, and acceptance criteria tied to business KPIs.

DevSecOps: Infrastructure-as-Code, CI/CD, policy-as-code, automated testing (unit, integration, synthetic).

DataOps: Medallion layers (Raw/Bronze, Curated/Silver, Serving/Gold), data contracts, and quality rules.

MLOps: Model registry, versioning, shadow testing, A/B evaluation, drift detection, and scheduled retraining.

4 Document Capture & AI Extraction Design

- Capture Channels: Bulk scan, mobile capture, MFD scan-to-email, secure inbox, and legacy backfile import.
- Image Standards: 300 DPI, lossless/compressed PDF/A-2b, grayscale/color per document class.

- Pre-Processing: De-skew, de-noise, auto-rotate, barcode/QR detection for batch separation, blank page removal.
- Classification: Heuristic + model-based classification to determine document type and route to correct extractor.
- Extraction Models: Prebuilt (invoices, receipts, IDs), layout (tables/lines), and custom neural models trained on labeled samples.
- Validation UI: Power Apps/React application for confidence review, field correction, reflow, and resubmission.
- Feedback Loop: Store labeled corrections to improve training datasets and raise precision/recall over time.

5 Data Platform & Analytics (Microsoft Fabric)

- Landing Zones: Bronze (raw JSON + OCR text), Silver (conformed entities), Gold (subject-area star schemas).
- Storage: OneLake Delta tables; partitioning by date/batch/document type for performance and cost efficiency.
- Transformation: Data Factory pipelines and Spark notebooks; quality checks and business rules enforcement.
- Semantic Models: Power BI semantic models with row-level security; curated datasets for self-service analytics.
- Operational Dashboards: Throughput, exception rates, SLA adherence, model accuracy, cost insights.

6 Search & Discovery

- Indexing: Azure AI Search indexes over extracted text and metadata; blob indexers with skillsets as needed.
- Enrichment: Language detection, key phrase extraction, PII redaction (optional), vector embeddings for semantic search.
- Query: Hybrid retrieval (keyword + vector) with semantic ranker; facets by document type, date, and status.
- Security: ACL/security trimming via application roles; support for delegated user context where required.
- UX: Standalone search portal, embedded in OpenText Smart View, and/or Microsoft 365 via webparts.

7 Document Management & Archiving (OpenText)

- System of Record: OpenText Content Server / Extended ECM for Microsoft 365 as authoritative repository.
- Metadata & Taxonomy: Synced from extraction; business workspaces per case/customer with templates and permissions.
- Records Management: Retention schedules, file plans, disposition reviews, legal holds, and audit trails.
- Archiving: Policy-driven tiering to Azure Blob (Hot/Cool/Archive); PDF/A normalization and checksum verification.
- Collaboration: Expose content in Teams/SharePoint via Extended ECM; co-authoring and workflow tasks.
- Integration: Connectors to ERP/CRM/case systems; content events to trigger downstream business processes.

8 Security, Governance & Compliance

- Identity & Access: Microsoft Entra ID RBAC/ABAC; managed identities for services; privileged access workflows.
- Network & Secrets: Private endpoints, VNET integration, Key Vault with customer-managed keys (where applicable).
- Governance: Microsoft Purview catalog, lineage, sensitivity labels; OpenText audit and records controls.
- Compliance: POPIA/GDPR alignment, ISO 27001 controls, evidence packs for audits, immutable logs.
- Monitoring: Centralized logs/metrics; alerts for SLA breaches; automated remediation for common failure modes.

9 Delivery Plan, Roles & Milestones

Milestone	Deliverables	Target Timeline	Acceptance Criteria
Phase 0 – Mobilize	Governance, environments, security baseline	Weeks 1–2	Environments ready; runbooks approved
Phase 1 – Discover & Design	Taxonomy, metadata model, solution design	Weeks 3–6	Design signed-off; backlog prioritized
Phase 2 – Build Platform	Azure/OpenText/Fabric/Search foundations	Weeks 7–12	Smoke tests passed; CI/CD operational
Phase 3 – Pilot Use Cases	3–5 document types live with validation	Weeks 13–20	≥95% availability; agreed accuracy targets
Phase 4 – Scale & Industrialize	Additional document classes; automation	Weeks 21–32	Throughput SLA met; ops dashboard live
Phase 5 – Operate & Optimize	Handover, training, benefits tracking	Ongoing	Run acceptance; KPI baseline achieved

9.1 Roles & Responsibilities

- Product Owner – Prioritizes use cases, defines acceptance criteria.
- Solution Architect – End-to-end architecture and NFRs.
- AI Engineer – Model selection, training, and evaluation.
- Data Engineer – Pipelines, lakehouse modeling, and quality rules.
- OpenText Specialist – Repository, records, and integration.
- Search Engineer – Index design, enrichment, and UX integration.
- Security & Compliance – Controls, evidence, and audits.
- SRE/Platform – Monitoring, reliability, and cost optimization.

10 Commercials

- A detailed bill of materials (BOM) and effort estimate will be finalized during discovery. Below is an indicative structure for budgeting purposes (licenses billed separately by vendors):
- Professional Services: Discovery & design, platform build, pilot use cases, scale-out, and change enablement.
- Azure Consumption: Document Intelligence, App Services/Functions, Storage, Fabric capacity, Azure AI Search.
- OpenText Licensing: Content Server/Extended ECM modules, records management, user seats (sized per role).
- Support & Managed Services: Run operations, model lifecycle, and platform optimization.

11 Risks & Mitigations

Risk	Likelihood	Impact	Mitigation
Variable scan quality reduces accuracy	Medium	High	Enforce scan profiles; pre-processing; feedback-driven model improvements
Model drift on new document layouts	Medium	Medium	Continuous monitoring; active learning and retraining cadence
Backfile conversion exceeds capacity	Medium	High	Phased backlog; burst capacity; third-party scanning partner
Security & compliance gaps	Low	High	Defense-in-depth; audits; privacy reviews and data minimization