Transforming Digital Archives

A Case Study on Seamless Migration and Enhanced Efficiency with Datablize DMS

Case Study: Digital Archive Migration at Anonymized National PSU

Executive Summary

A significant Indian Public Sector Undertaking (PSU) undertook a comprehensive modernization of its enterprise archive, successfully migrating over 10 million historical and active documents into the Datablize Document Management System (DMS). This transformation provided a centralized repository, enhanced search capabilities, automated information extraction, and ensured compliance with retention policies, all while maintaining seamless business operations. Key outcomes included a **37% reduction in storage** requirements, a **search latency of less than 1.2 seconds** at the 95th percentile, and a **60% improvement in document retrieval times** for case handlers.

Business Context & Drivers

Challenges

- Dispersed content across outdated document management systems, file shares, and email storage.
- Slow document retrieval affecting audits, Right to Information (RTI) requests, and regulatory submissions.
- Inconsistent metadata, prevalent duplicate documents, and unclear version histories.
- Need for **compliance-grade governance** including retention, legal hold, and audit trails.
- Requirement for secure remote and field access incorporating Single Sign-On (SSO) and Multi-Factor Authentication (MFA).

Objectives

1. Consolidate archives into a secure, scalable DMS ensuring **zero data loss**.
2. Standardize taxonomy and metadata to enhance searchability and reporting.

3. Automate document classification and Optical Character Recognition (OCR) for increased accessibility.
4. Implement lifecycle governance, including **retention policies, legal holds, and auditability**.
5. Ensure migration is conducted with **no business downtime** and maintains data integrity.

Solution Overview

Technology Stack

- **Platform:** Datablize DMS with Intelligent Document Processing (IDP) and Workflow Business Process Management (BPM).
- **Hosting:** Private cloud with multi-availability zone high availability; SSO through OpenID Connect; MFA for privileged roles.
- **Content Types:** Includes policies, contracts, invoices, engineering drawings, human resources files, correspondence, and project dossiers.
- **Interfaces:** Web application, APIs for ingestion and metadata, watch-folders, and mail-integration.

Key Capabilities

- AI-assisted **auto-classification** and key-field extraction, supported by human-in-the-loop validation.
- **OCR normalization** for various formats, supporting English and Indic scripts.
- **De-duplication** using SHA-256 hashing and near-duplicate checks.
- **Taxonomy and metadata standardization** with controlled vocabularies and picklists.
- **Records management** including retention schedules, legal holds, and immutable audit logs.
- Advanced search capabilities including full-text, semantic, fuzzy, proximity, and Boolean queries with hit-highlighting.

Migration Approach & Timeline

Phase 1 – Discover (Weeks 1–2)

- Conduct inventory of source documents; sample 3% of files for quality and format analysis.
- Develop risk register, success metrics, and stakeholder engagement plans.

Phase 2 – Design (Weeks 3–4)

- Create target taxonomy, content types, and define mandatory fields.
- Design security model (RBAC/ABAC), foldering, and retention schedules.
- Map source to target metadata, aligning with controlled vocabularies.

Phase 3 – Pilot (Weeks 5–6)

- Migrate 250,000 documents from four departments; measure classification precision and recall.
- Validate OCR accuracy and set extraction confidence thresholds.

Phase 4 – Bulk Migration (Weeks 7–12)

- Utilize parallelized movers with eight worker pools; verify transfers with checksums.
- Detect near-duplicates using text-fingerprinting; quarantine for review.
- Continuous indexing; implement incremental cutover by department.

Phase 5 – Cutover (Week 13)

- Transition legacy systems to read-only; perform delta synchronization and final integrity audit.

Phase 6 – Optimize (Weeks 14–16)

- Tune re-indexing; implement search synonyms; deploy dashboards/KPIs; provide training and handover.

Technical Design Highlights

Ingestion & Validation

- Connectors for legacy DMS export, SMB shares, email PST/IMAP, and multifunction printer watch-folders.
- **Checksum chain:** Verify SHA-256 at source, during transit, and at target; auto-retry on mismatch.
- Normalize files to PDF/A where possible; clean page images for OCR.

Classification & OCR

- Utilize hybrid rules and machine learning models; route based on confidence to validation queues.
- OCR pipeline supports multiple pages, skew/deskew corrections, noise reduction, and table extraction.

De-duplication & Near-dup

- Identify exact duplicates through cryptographic hashes; near-duplicates through text-similarity and image perceptual hashing (pHash).
- Merge policy retains **earliest provenance**; links later versions to lineage.

Security & Compliance

- Implement SSO (OIDC/SAML), **MFA** for administrators, and fine-grained RBAC/ABAC.
- Enable at-rest encryption and secure TLS 1.2+ in-transit; ensure immutable audit log exports.
- Bind retention and legal hold policies to content types and event triggers.

High Availability & Scale

- Employ stateless workers with queue-based backpressure; utilize multi-availability zone object storage.
- Conduct blue-green deployments and automated failover drills quarterly.

Change Management

- Establish a network of champions across departments; provide role-based training.
- Offer contextual help, quick-reference guides, and video micro-lessons.
- Deliver hypercare for four weeks post-cutover; conduct bi-weekly office hours thereafter.

Quality, Controls & KPIs

Migration Quality Gates

- Ensure ≥ 99.95% checksum match rate; send unresolved < 0.05% to remediation.
- Implement sampling protocol: 0.5% random and 100% of high-risk classes manually verified.

Search & Retrieval

- Achieve 95th-percentile search latency of **< 1.2 seconds** post-tuning.
- Improve first-result usefulness by **48%** compared to baseline, as measured by click-through rates.

IDP Accuracy (Pilot → Steady-state)

- Achieve auto-classification F1 score of **0.90 → 0.94** with feedback learning.
- Maintain field-level extraction accuracy for fixed forms at **≥ 98%** with 0.85 confidence threshold.

Operational Throughput

- Average ingestion rate of **180,000 documents per day**; peak day handling **450,000+ documents**.
- Ensure index freshness of **< 60 seconds** from write to searchable (95th percentile).

Results & Impact

- **Unified archive:** Establish a single source of truth across nine regions and 14 departments.
- **Storage optimization:** Achieve a **~37%** reduction through de-duplication and compression.
- **Faster work:** Realize a **~60%** reduction in average retrieval time for case files.
- **Audit-ready:** Ensure end-to-end lineage, immutable logs, and automated retention.
- **Lower risk:** Enhance legal hold and access governance to reduce accidental exposure.

“The new archive lets our teams find and act on information in seconds, not minutes—while checksums and audit trails keep us compliant.” — Deputy GM, Records (anonymized)

Before vs. After

Dimension	Before (Legacy)	After (Datablize DMS)
Repositories	6+ disparate stores	Unified content hub
Search	File-name only; slow	Full-text + semantic; p95 < 1.2s
Metadata	Inconsistent; missing	Standardized taxonomy & required fields
Duplicates	Rampant	Hash-based de-dup + lineage
Governance	Ad-hoc	Policy-based retention & legal hold
Audit	Limited logs	Immutable, exportable audit trails

Lessons Learned

- Invest early in **taxonomy design**—it yields significant long-term benefits.
- Conduct pilot programs with **representative data**; fine-tune confidence thresholds before scaling.
- Treat de-duplication as an **ongoing hygiene** practice, not a one-time task.
- Align cutover with business cycles to minimize change fatigue.

Next Steps

- Expand IDP templates to encompass additional document classes.
- Integrate e-signature for approvals and enable process mining dashboards.
- Establish quarterly governance reviews and search-quality tuning.

At a Glance

- **Volume:** 10M+ docs (≈ 12 TB)
- **Timeline:** 16 weeks to cutover
- **Peak Throughput:** 450k+ docs/day
- **Search p95:** < 1.2s
- **Storage Optimization:** ~37%
- **Retrieval Time:** ~60% faster