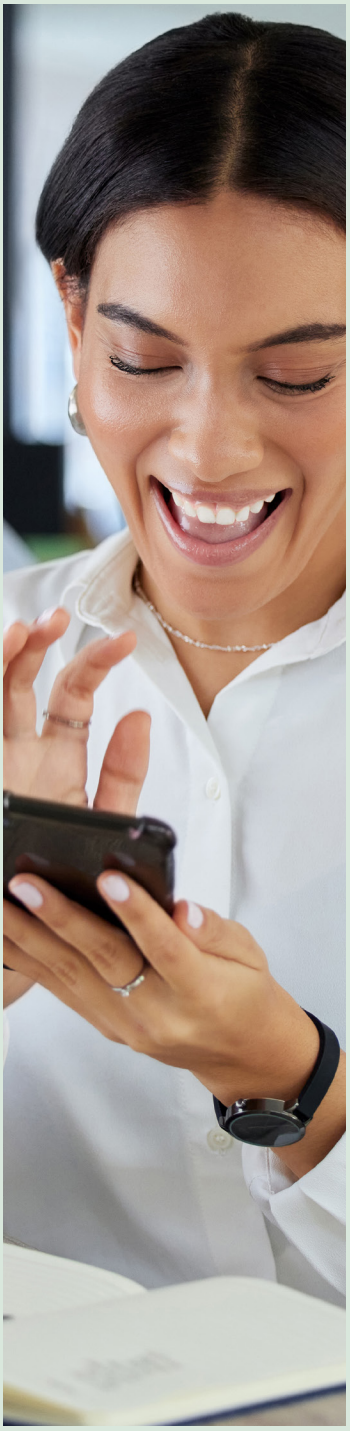
WHITE PAPER



DOCS OPEN/EDOCS TO FILENET JOURNEY

Infosys®
Navigate your next

Executive Summary

Moving from a legacy ECM application to a newer ECM stack is a difficult and complex, but necessary step for every IT department. While out of support application stack, ageing hardware & performance challenges pose risks to business continuity, regulatory & compliance requirements also add to the equation, aided by the desire of IT departments to move towards a strategic solution, with access to latest digital features. The task of migration is often complicated by limited infrastructure capabilities of legacy applications, further exacerbated by data/application characteristics challenges (e.g., proprietary data formats, data quality, lack of vendor support), necessitating careful evaluation of migration and modernization options & alignment with the target state.

Moving from legacy DocsOpen or eDocs stack to a newer ECM platform is one such scenario. This whitepaper narrates the journey with FileNet based stack chosen as the target platform. The paper discusses moving from legacy DocsOpen/eDocs stack (DocsOpen 4.x/eDocs DM 5.x) to FileNet 5.5.x & Datacap 9.1.x based stack. It provides a view of source & target application stacks, factors affecting migration, options to be considered, and typical issues associated with data and functionality migration. Note that the functionality migration scope covered in this paper is core document management (i.e., capture, ingest, search/retrieval/update) only

In the interest of brevity, the remainder of this paper refers to DocsOpen version 4.x/eDocs version 5.x as DocsOpen/eDocs only, and it should not be confused with later versions of eDocs products

Overview

Legacy DocsOpen (earlier PCDOCS) & its successor eDocs 5.x applications, were quite popular in the legal fraternity, with wide adoption in law firms & legal teams of large enterprises across the world. Typical use cases were communications data ingestion/search using mail client integration (e.g., Outlook) and legal data ingestion/search/update using desktop applications (office applications, Windows explorer, DM thick clients etc.). The experience was further enhanced using custom thick clients, e.g., 3rd party application integration, realization of client-matter functionalities using custom client features like create/update/delete/lookup etc.

DocsOpen/eDocs' Imaging thick client capabilities include ability to search/retrieve/update documents directly from the repository and add annotations, apart from core capture functionalities. In the finance sector, this setup had a sizable footprint, usually with a capture based ingest & thick/web client (Webtop) based search/retrieval setup.

IBM FileNet stack is one of the leading players in the ECM landscape with a wide presence across industries, providing enterprise content management capabilities to add, access, and manage different types of content, making it available to users as required, with the ability to further expand the capabilities using add-ons. Coupled with IBM DataCap, it offers an extensive set of features to handle the data journey in enterprise, starting with capture/ingest, indexing, storage, versioning & search/retrieval. User experience is driven by ICN (IBM Content Navigator) web client, allowing plug-ins/ extensions to augment business functionalities further, e.g., web based capture, Office integration for desktop Office capabilities, EDS plug-in to integrate data from sources other than the repository etc.

Product Ecosystem & Components

Legacy DocsOpen/eDocs DM resided in Windows ecosystem (no Linux support), using COM based DM API layer, supporting COM-enabled languages (VB, .Net stack etc.) & IIS web server for Webtop & DM API deployment. It followed the classic three-tiered architecture, with RDBMS & file share for persistence, DM server & DM API for processing layer and thick clients/DM extensions (add-on modules) and/or Webtop for user experience layer. DM APIs facilitated consumption of DM server functionalities for OOB clients (Webtop, Imaging desktop etc.), as well as further extension of core functionalities, e.g. developing add-ons, custom forms, e-mail Integration modules etc. For capture, Imaging desktop, a thick client tightly coupled with DM server, provided client-only capture capabilities (scan documents, add annotations etc.). It also allowed real time data search/retrieval/update from the data repository.

While this product stack was one of the leading players in its heydays, the legacy nature of the product limits the technical & functional options available to cater to today's business requirements

FileNet stack supports a number of platforms, including multiple Linux flavors, Windows as well as AIX, with on-premise, cloud & containerized implementations. For user experience layer, ICN is used, with CPE performing the processing layer's heavy lifting & RDBMS & file share for persistence layer. FileNet follows a Java EE based application model, with CPE deployed inside the EJB layer of a Java EE application server (e.g., WebSphere). A set of Java APIs expose the CPE functionality to OOB (ICN, ACCE etc.) & custom clients (.NET API stack is available too). The SOAP based web services allow web based consumption of CPE functionalities.

Datacap is the enterprise capture solution, offering full range of automated capture functionalities and seamless integration with FileNet.

ICN runs in a web container on a Java EE application server (e.g., WebSphere), using plug-ins to implement various feature add-ons, and Deaja viewer to facilitate server based viewing, allowing a lightweight but secure end-user experience, requiring only a supported browser. ICN for Microsoft Office is an end-user .Net based component integrated with ICN, providing the office products based data access from user machine.

Datacap Navigator (an ICN plug-in) provides distributed capture capabilities to end users, via web UI with Web Twain drivers at user end (FastDoc provides the thick client alternative). Datacap supports Windows platform, with processing power provided by taskmaster server components, persistence by RDBMS & file store & web capabilities using IIS based web server.

The vast set of digital features & functionalities available, with a large product support matrix as detailed above, make FileNet a good choice to serve the business needs of the enterprise.

Migration Considerations

For the 'What' part of the migration, i.e., planning the move from legacy DocsOpen/eDocs stack to a FileNet based stack, the following factors should be given due consideration –

- **Source State** – The source state drives the extraction solution, i.e., product based or custom approach. If the source has the product API stack installed (installed separately from the core product), with enough compute in place, then extraction through API should be considered. If not, then an inorganic approach would work better (e.g., extract data from database & file share using scripts, bypassing product APIs)
- **Business Impact** – If the source system is live, with consistently heavy user load, then the API based data extraction window needs to be planned carefully. This can prolong the overall migration duration for large data volumes, as extraction is usually the slowest part of the ETL cycle. It is advantageous to use database & storage based extraction in this scenario as compute & processing threads can be increased as per the need, without any impact on live applications.
- **Upstream Feed Redirection** – Upstream feed (i.e. feeds from other systems with ECM as destination) re-direction to the new solution needs to account for how (redirection options), where (underlying infrastructure changes) and what (upstream dependencies & unknowns) factors. The batch feed processing can be handled using a CEBI/custom utility based consumption in FileNet, often accompanied by underlying infrastructure changes as well, e.g., use of an SFTP based transfer, with enterprise scheduler for batch automation (e.g., Control M). Other factors impact this activity significantly as well, e.g., dependency on upstream/intermediary applications, lack of SME knowledge around legacy feed processing logic etc.
- **Migration Infrastructure** – Following are the important factors when it comes to migration infrastructure planning –
  - Compute capacity planning – Having enough compute & storage space is critical as the migration environment does the heavy lifting for the overall migration process.
  - Number of environments – Ensuring a sufficient number of migration environments, to iron out the migration scenarios, QA & performance aspects.
  - Migration tools – Firming up the infra requirements for the tool e.g. storage type (block storage), database, tool UI & add-on requirements.
  - Antivirus scanning – Suppresing Antivirus & background processes, as these processes can often consume significant compute power, slowing down the overall migration.
  - Compliance – Conducting compliance requirements evaluation, for obtaining the required exceptions where necessary (for storing the user data in migration environments), to avoid the need for additional infrastructure elements (e.g., tools for data masking/redaction).
- **Target State Expectations** – for data movement, it is beneficial to consider the target data state requirements as an input to migration approaches. For example, a scenario often encountered is 'migrate everything', moving massive data volumes to target, only to purge legacy data based on records management later. The duration of migration is directly proportional to the volume of migration; hence this prolongs the overall schedule.

  Another scenario often playing out, is to build complex solutions for migrating proprietary legacy data elements e.g., annotations/ document linkages when migrating them as indexes/individual documents respectively, can serve the business purpose.

  Similarly, any format transformation requirements should be carefully planned for & considered (e.g., legacy formats' conversion to pdfs)
- **Delta Characteristics** – Delta data (new data generated in source, post migration) too, can throw some unique challenges to deal with. For example, new document version(s) and/or annotations added post main migration run in source DocsOpen/eDocs data set, can make the delta migration quite tricky.

  For heavy activity applications, the delta volume too, could be quite high (seen frequently in finance organizations), requiring multiple deltas to be migrated. The migration approach needs to plan for such scenarios.

Migration Approach

The 'How to' part of data migration can take different routes, depending upon many variables as explained below -

Data Extraction - As detailed earlier, the data extraction from legacy DocsOpen/eDocs can be done using either the API based approach, or storage layer based approach, as detailed below –

- **API based** – Where API based extraction is deemed feasible, custom .Net scripts can be developed to extract data using DM API (using PCDClient type library). While Java is technically supported for DM API consumption, its usually tricky since it would require COM wrappers to be available to access the DM API.

Data extracted using API is fed to an interim migration environment, with transformation scripts mapping source indexes to target data model & moving content to file share.

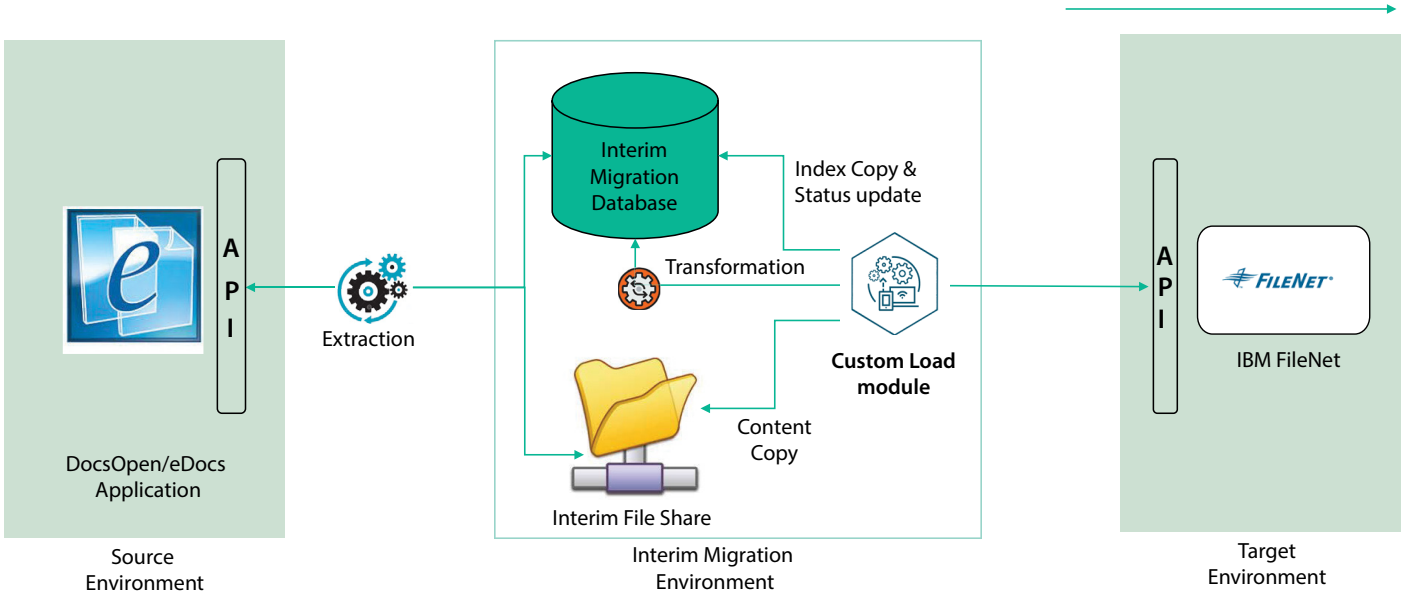


Figure 1: API based extraction (and migration) approach

- **Database & File Share based** – Often, the legacy infra/ application compute limitations or unavailability of API layer itself, necessitates use of alternate approaches. In such cases, a database & file share based extraction approach works best. In this, the legacy application data is moved to an intermediate migration environment. Source file share copy is mounted to a migration server & the source database backup is restored to a migration database, providing the business context of ECM files. This interim migration environment is then used as the data source.

  While DocsOpen/eDocs has an extensive RDBMS schema, it is possible to reconstruct the source data model using few key tables. The most key table is the Profile table, containing all document indexes, supported by the Versions table containing data on all document versions. Component table contains data on all document component files & Path containing the file names. For specific industries (e.g., Finance/Legal) additional tables contain the specific data that need to be extracted.

  In this setup, custom scripts extract document properties from interim migration database's source tables & move to target table(s) in an interim migration database, based on source to target mapping.

  Source content files are stored in human readable format, and hence doesn't require any special extraction considerations.

  Annotations (where present) require some planning & few options to be considered based on target state (discussed in the next section).

  Below diagram provides an overview of this process -

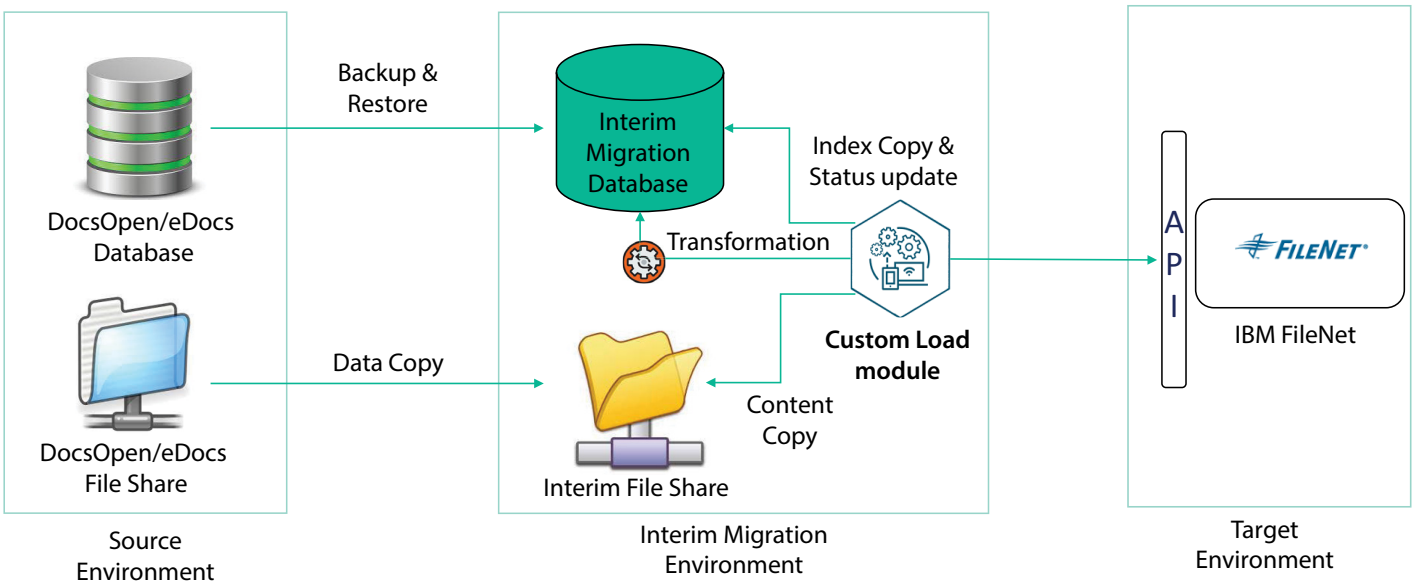


Figure 2: Database & file share based extraction (and migration) approach

Data Transformation – For DocsOpen/eDocs to FileNet migration, the data transformation techniques revolve around the standard ECM practices, mapping source document properties to target data model, as well as content transformation where applicable. Following are few key factors that should be taken into account –

- Source to target mapping must be created early in the program & agreed with stakeholders to avoid any challenges at later stages. Changes at later stages can incur considerable re-work & result in cost & schedule impact.
- Target data model should consider the source data characteristics (format, length, runtime behavior, structure etc.) & clear rules must be available for exceptions.
- Transformation output should be validated thoroughly with legacy source data set (e.g., handling annotations on multiple versions). For example, at times, >100 versions of a given document can be found in legacy DocsOpen/ eDocs applications, with heavy annotations. Target data model should be able to accommodate such extreme cases.
- New data elements generated as results of transformation (e.g., text annotations export to properties) should be thoroughly verified for data integrity.

Below diagram provides a sample source to target mapping combination –

Source				
Legacy Property symbolic name	Type	Length	DB Column	Rule
F_docId	String	15	A12	NA
F_document type	String	15	A15	NA
F_accnumber	String	20	A18	NA
….	….	….		

Target					
Target document class	Symbolic name	Property display name	Property symbolic name	Type	Length
Loans	LoanDocs	Legacy Doc Id	LegacyDocId_S	String	20
		Document Type	DocType_S	String	15
		Account Number	AccNumber_S	String	20
		…	…	…	

Figure3: Source to target mapping sample

Data Upload – While FileNet provides bulk tools for data uploading, usually custom tools provide more flexibility for loading the migration data. Here are a few points to consider for uploading data –

- For data upload, a java module using CE java APIs can be used to load the data into FileNet. The module should run multiple instances to save time and maintain logs to ease debugging of the issue.
- Load scripts can write back target load status in interim table (e.g., load success/failure against the legacy data row, to indicate the load status) to provide capability to easily track load failures/issues (as depicted in figure 1 & 2, earlier in the section).
- Reconciliation reports can be used to validate the data migrated, to provide counts based as well as individual document based reconciliation.
- Sample based QA should be performed against the migrated Prod data, to ensure data quality.

Additional Considerations

Besides the standard ECM best practices, below factors should be considered for eDocs/DocsOpen functionality movement to FileNet stack –

Access Flexibility

Think of Access Anywhere, empowering users to perform business functions from a browser, with no/low component prerequisites. While users would usually prefer functionality similar to source DocsOpen/eDocs applications, use FileNet's browser based features instead, where possible, to allow machine independent access, lower costs & reduce complexity (no desktop package rollouts for thick clients) as well as significant maintenance reduction, thus improving ROI.

As an example, FileNet entry template based drag & drop, with search templates for search/retrieval functionality can closely match the Legacy mail thick client's ingest/search functionalities.

Similarly, a web based capture process (Datacap navigator) can ease the package/fixes roll out hassles typical for the capture thick clients.

User Training

While a good user training regime is essential for any switchover, the effects can be more pronounced in certain scenarios especially the movement from thick clients to browser based UIs. The change management process should emphasize on such cases in user education, with a comprehensive instruction set (e.g., demo, user guides).

For example, the client-only capture nature of legacy imaging desktop allows user to shuffle/delete pages at will, however Datacap navigator doesn't allow main page (1st page of the scanned document) deletion or selecting any page before the main page during indexing.

Status Quo Changes

At times, legacy application remediations lean towards moving current feature set to target as-is. While this is dependent upon business requirements, options to consolidate/retire functionalities should be explored & business should be encouraged towards logical outcomes rather than just moving the current set as-is.

An often found example of this is legacy custom UIs where the underlying data set is past its active life, with very infrequent access. For such scenarios, putting a BAU support process in place, is a better option than building UI features to achieve it.

Another frequent scenario is ad-hoc features & security setup built over legacy application's lifetime, which can be consolidated & simplified.

Automation

Legacy DocsOpen/eDocs setups relied a lot on manual processes or customizations to perform data ingest/indexing/validations. FileNet & Datacap's vast set of automation capabilities allow automating many such scenarios & should be used where the opportunity presents itself.

For example, Datacap can pull in data from a given mailbox, validate, index & ingest to FileNet & send notifications to a specified set of users without any manual intervention, allowing quick to market solution & brining in process efficiencies.

Key Challenges

Moving from legacy DocsOpen/eDocs stack has its own set of challenges, along with typical migration issues. Below are the few issues often encountered –

- **Annotations** – The legacy imaging desktop client created the image annotations in a proprietary format (Wang/Kodak), storing the annotation data inside tiff image header. Modern desktop clients (image viewers, browsers) & server based viewers (Daeja viewer) don't support this format.

  To extract the annotations, a custom module can be built using –

  - either a proprietary API (e.g., GdPicture), to burn the annotations inside the tiff image
  - or an open source tiff API (e.g., Apache Imaging), to extract annotation text from tiff header & store it as a property
- **Vendor Support** – At times, there are unknowns present in legacy application functionalities or complications related to data extraction (e.g., annotations). Ensuring product vendor support beforehand is thus beneficial, helping resolve any source analysis/data extraction related issues.

  If no vendor support agreement is in place, such issues may require custom solutions & may significantly increase the risk & cost of the program.
- **SME Challenges** – Availability of SME with source system knowledge can be a differentiating factor for legacy DocsOpen/eDocs migration. These are usually 1.5 decades or older setups, thus may have very few resources with end to end system knowledge (e.g., custom application logic, data processing rules). Unavailability/limited availability of knowledgeable SME can pose serious challenges to migration & should be planned for.
- **Test Data Availability** – Since the source DocsOpen/eDocs setup is likely from early/mid 2000s, non-prod environments often don't have the complete representative test data for all scenarios. In such cases, option to test with production data should be explored.
- **Data Quality Issues** – While not exclusive to DocsOpen/eDocs stack, legacy data quality issues are encountered often and need clear rules to be formed & agreed with stakeholders. Some of the issues typically seen are missing/wrong mime types, corrupted (total/partial) content files, 0 KB files & junk/missing metadata values.

Summary

Moving from legacy eDocs/DocsOpen solution to a FileNet & Datacap based solution may seem challenging & overwhelming, however with detailed analysis of source data & features, a thorough migration plan, coupled with a clear pathway to handle the typical pitfalls & user training can help ease this journey.

Infosys, with its vast set of skilled ECM resources, mature migration tools & processes, combined with migration expertise acquired across industry verticals can make the transition easier, by bringing in a well-planned migration approach, to fit in a FileNet & Datacap based solution as a perfect replacement for legacy stack and help realize valuable benefits for organization. It can help improve the user experience journey, de-risk the organization's data, ensure compliance as well as bring in process efficiencies and improve ROI.

About the Author

Ravi

Senior Technology Architect with Infosys

He is working primarily in the OpenText & IBM ECM landscape. When not solving design challenges, he can often be found immersed in automobile materials

For more information, contact askus@infosys.com

Infosys®
Navigate your next