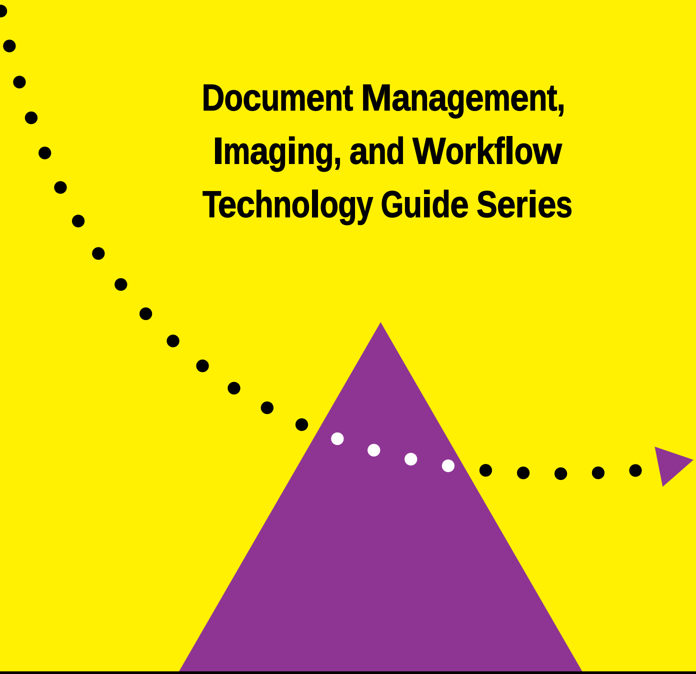
# Document Warehousing:

## Building an Enterprise Document Repository



Document Management, Imaging, and Workflow Technology Guide Series

This guide has been sponsored by

Kodak

# Kodak Business Imaging Systems: Enabling Better Document Access

Kodak Business Imaging Systems makes it easier and more cost-effective for business and government customers to file, find and distribute information—in the form of document images—with complete assurance of future access.

Kodak first focused its imaging expertise on the problems of business in a pioneering 1928 application. When a banker asked for a secure, compact way to record transactions, Kodak invented an automatic camera that microfilmed checks at production speeds.

In the years since, Kodak has responded to business needs with a full range of imaging capabilities that span media, hardware, systems and solutions, and services. These help customers improve their document imaging by addressing issues of access, space requirements, administrative cost, cycle times and quality of service. Today, customers at over 20,000 installations in the Americas, Europe and the Pacific Rim depend on Kodak to facilitate their document-intensive processes. The diverse industries involved include financial services, insurance, transportation, manufacturing, health care, education and service providers, among others.

## From image capture to archival storage—and every point in between

Kodak offers problem-solving capabilities for virtually every stage of the document life cycle. Some are components, others are complete systems. Kodak is committed to media independence and open architecture in the design of its products. As a result, we increasingly are becoming integrated into our customers' overall information access as part of a "document warehousing" strategy.

**Media: Choices for Challenges.** Kodak offers what is perhaps the world's widest range of media, including film, writable CD and optical disk. With this selection, customers can choose the medium that's most appropriate for their specific applications. For example, Kodak microfilms provide legal, archival storage at an extremely low cost. Images may be retrieved, scanned and sent to printers, fax machines and networks using Kodak digital workstations. Likewise, Kodak automated CD jukeboxes and OD libraries provide immediate access to large volumes of digitally-stored documents for workgroups and departments.

**Hardware: Components with Staying Power.** "Industrial strength" is the phrase often used by customers to describe the hardware from Kodak used to capture, store, manage and distribute document images. To convert paper documents to images, Kodak supplies high-speed production and medium-volume document scanners and microimagers. Kodak scanner-microimagers simultaneously create a digital image for on-line use and a microfilm image for archival storage in a single-step process. Later in the document life cycle, the award-winning *Kodak Digital Science*™ Document Archive Writer "writes" digital document images out to film in analog form. Virtually obsolescence-proof, this technology meets the requirements of applications that must retain records to meet legal requirements or because a document remains active over the lifetime of a person or property.

(continued on inside back cover)

EASTMAN KODAK COMPANY
Business Imaging Systems
Corporate Headquarters: Rochester, New York 14653-6304
(800) 243-8811

In other areas contact the Kodak company serving your country.

Visit our home page on the World Wide Web at
http://www.kodak.com

"The significant problems we face
cannot be solved by the same level
of thinking that created them."

## Table of Contents

# What is Document Warehousing and Why is it Important?

As information technology has evolved, new computing platforms, relational database models, telecommuting and other information technologies have enabled organizations to fundamentally change the way they conduct their operations. However, these technologies have done little to automate the document-centric business operations. Over 90% of the information in today's offices exists in document form, or as "unstructured information". Comprised of a diverse collection of handwritten, printed and desktop-created documents, such as reports, forms and correspondence, as well as voice transcriptions, art work and photographs, companies are now grappling with storing and managing the information resident in these diverse information bases.

To that end, companies have implemented technological solutions appropriate for each type of information: document imaging systems to handle digitized paper documents, document management systems to control electronically-generated documents, computer output to laser disk (COLD) systems to process mainframe-generated reports, workflow applications to route and process specific work items that contain documents, and groupware applications to facilitate the collaborative sharing of information and completion of projects. The most recent technological solution is the use of web servers and browsers to create web-enabled applications that publish and distribute information across the entire enterprise.

But as these individual technologies have evolved, each has developed proprietary and unique models for document access, retrieval and security. With no common storage or inquiry method, it is impossible for organizations to access the shared information base that is represented by this collection of document systems. That's where the concept of document warehousing comes in. By grouping individual information bases into a repository with a single capture, access, retrieval and security model, users throughout the organization can obtain, communicate and distribute the combined information assets of the corporation.

The evolution of document warehousing has it roots in data warehousing, which essentially creates a centralized repository of the structured information contained within separate and often isolated databases. By integrating an organization's multiple, related data stores, data warehousing provides business managers throughout the organization with rapid and efficient access to information for analytical and decision support purposes. With document warehousing, multiple unstructured information bases are stored in an enterprise document repository (EDR), which enables rapid, efficient and universal access to information not represented by existing data warehousing models. By providing access to unstructured business information, the document warehousing solution provided by an EDR delivers a significant and compelling advantage to an organization.

Document warehousing is not a single product or technology, but a standards-based environment that enables users to capture, link and retrieve various types of documents in a form that can be easily accessed and immediately distributed. Nor is document warehousing available in a packaged form today. Instead, standards and products are evolving to support the level of interoperability required to create an EDR. And, some technology suppliers are going one step further: building the software and systems that will make document warehousing a reality.

This technology guide explores the concept and benefits of document warehousing, discusses what the functional capabilities of an EDR must be, outlines potential architectural considerations for an EDR, and finally, provides guidelines for evaluating, preparing and implementing a document warehousing strategy within your organization.

## Enterprise Information Challenges

Today's organizations are at once drowning in and starved for information. The corporate computing revolution that has occurred during the last several decades has dramatically enhanced our ability to capture, compile, report and, most of all, create "data". With virtually every transaction recorded in databases, the growth in structured data has exploded to such a point that users are looking for new technology solutions that enable them to rapidly and accurately access the information they need among the terabytes of structured data available.

The growth in unstructured data that does not fit into the "row and column" database paradigm is just as overwhelming: Tens of millions of users, who have been empowered with personal computers, are each generating hundreds and thousands of documents per year, including contracts, letters, memos, reports, spreadsheets, invoices, expense reports and sales reports. Each day, 2.7 billion new sheets of paper are generated by U.S. business workers. In fact, our ability to create data now exceeds our ability to extract and utilize the information resident in that data.

The exponential growth in both structured and unstructured data bases poses critical challenges to corporations: how to access, control and leverage the information resident in these myriad data sources. An organization's competitiveness is increasingly linked to its ability to manage and distribute these information knowledge bases. Whether the business objective is improved customer service, getting products to market faster, or reducing operational costs, sharing and managing all relevant information is a crucial imperative.

The technology of data warehousing is aimed at solving that challenge for structured data management. Data warehousing creates a centralized repository of information, which is built from separate and often isolated database systems. Accompanied by tools for analysis and rapid application development, data warehousing technology allows business managers throughout the organization to access, review and act on this collection of data.

There is no comparable approach for the management of document or unstructured information bases, which collectively represent the overwhelming majority of information within an organization. Instead, in an attempt to gain control over the various types of unstructured data, such as electronic documents and spreadsheets, color and black-and-white photographs, mainframe-generated billing or customer reports, or digitized paper documents, organizations have implemented individual, discrete solutions based on the following technologies:

- Document management

Designed to provide more control and better management of computer-generated data files (especially word processing documents), document management technology adds enhanced file security, revision control, file descriptions, extended file names and user access privileges to the basic file directory management features of the computer operating system.

- Document imaging

Document imaging technology enables the input, indexing, management, storage and retrieval of digital image files, which typically originated as scanned or faxed forms and letters. These image files require specialized viewing software, compression and decompression software, drivers for specialized subsystems, and support for a range of specialized storage environments.

- COLD

Computer output to laser disk technology allows the processing, indexing and storage of computer-generated, formatted reports on computer media. Previously, this data would be output to printed report pages and/or to microfiche. With COLD, users can electronically search, view, print and process the information contained in any report.

- Workflow

With workflow technology organizations can define work processes in terms of participants, inputs, outputs and work flows, and automatically route work tasks and the information required to perform those tasks throughout the organization.

Other workflow features include the ability to set rules and policies that govern the flow and fulfillment of work tasks, the ability to monitor workloads and reallocate resources accordingly, and the capability to revise the flow of work after identifying inefficiencies or bottlenecks.

- Groupware

Groupware technology permits the collaborative sharing of work and information among individuals and groups. Based on a messaging infrastructure and a document database, groupware provides coordination and collaboration features for group projects and activities.

- Internet/Intranet Web Sites

Companies are aggressively exploiting the platform-independent nature of the Internet to better manage the publication and distribution of unstructured data across the enterprise. Private Web sites make it easy for users throughout the organization to access information created with different applications. In addition, organizations are now beginning to web-enable groupware and workflow applications for collaborative work management.

While each of these technologies offers compelling benefits, there has been no integrated and universal approach to obtain related information across multiple information bases and document stores. As the individual technologies have evolved, each has developed unique and often proprietary access, retrieval and security approaches that make it impossible to sit at a single workstation and, at one time, access all the information relevant to a specific account, customer or project. In essence, each system manages information assets that are generally inaccessible to the enterprise.

Additionally, with the exception of some archival document imaging solutions, each of these technologies is almost exclusively focused on the most active stages of the document life cycle: when unstructured data is created, stored, retrieved, routed, edited, processed and discharged. During this stage, documents are typically retrieved often and used by several individuals within a workgroup or department. Thus, the system must be optimized to retrieve information as quickly as possible, and be capable of routing documents to appropriate individuals for processing.

Eventually, the document enters a more latent phase in its life cycle, a phase characterized more by archival and storage characteristics than by processing characteristics. At this stage, the focus turns to long-term storage and information asset protection. Users need to store large volumes of documents economically and safely, providing protection from technology obsolescence. While retrievals do not have to be immediate, they do need to be efficient and available to a broad population of users (see Figure One).

Few of the technologies described above emphasize retention-related characteristics, such as the economic storage and retrieval of large volumes of documents, obsolescence-protection, long-term enterprise access to documents, and secure long-term storage for documents as they enter their latent stages. Most of today's document management and workflow vendors offer little in the way of storage management, in effect using the file system and drive-letter mapping to store document files. Those systems that do address document storage do so by interfacing with hierarchical storage management software (HSM), which is primarily designed for data backup and not well-suited to service document retrieval requests from potentially thousands of users over wide areas.

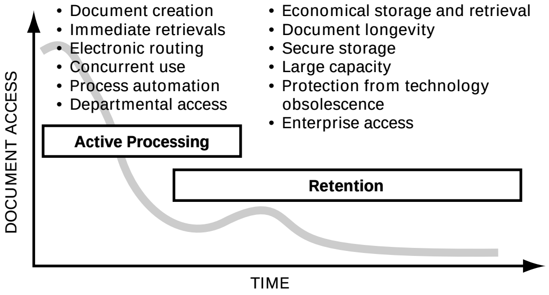


**Figure 1: The Document Lifcycle**

## The Solution: Warehousing Documents in an Enterprise Document Repository

Just as data warehousing has evolved to address the multitude of structured data repositories created by line-of-business transaction systems, document warehousing can provide a universal and consistent access, storage and retrieval model for the disparate document repositories contained within proprietary imaging, document management, workflow, output management, groupware and web-based systems. With document warehousing, users can aggregate and store a variety of unstructured information bases in an EDR.

An EDR supports the capture, indexing, archive storage and retrieval of large volumes of unstructured data or document information over long periods of time. The EDR allows companies to implement a centralized or distributed document warehousing strategy that pulls business documents from multiple sources, places them within a common storage and management infrastructure, and makes them available to users throughout the enterprise.

The EDR is uniquely positioned to enable businesses to leverage the value contained within their document-based information resources. It begins by allowing them to capture documents from locations that are currently difficult to access within proprietary document processing systems. These documents, which exist in multiple formats with variable content and metadata, are then transformed to a standard storage format and bound to a set of attributes consistent with the document indexing standards of the enterprise. The documents are then archived to the most appropriate storage media considering both corporate policies for long-term retention and subsequent user needs for network access. Once archived, the documents are available for access over local and wide area networks using standard desktop or web-based client software for repository browsing, search and retrieval.

## The Benefits of the EDR

Document warehousing is important for the same business reasons data warehousing has become an essential part of the corporate information technology landscape: it allows companies to leverage their corporate information assets to improve customer service and achieve a competitive advantage.

An EDR offers tangible and compelling benefits to users, including:

- a consistent, single point of access to document (unstructured data) information stored in different systems. This enables users throughout the enterprise to access integrated information bases for cross-functional and multiple business purposes.

- the ability to provide long-term and very high-capacity storage. Rather than rely on disparate storage models for various unstructured document systems, an organization can establish an enterprise-wide storage strategy that is secure and disaster-survivable, and that can deliver required document information on demand. By eliminating customer dependency on multiple vendors for archive and storage management, organizations can also gain economies of scale for unstructured data storage and archiving.
- the opportunity to establish a rational and consistent enterprise-wide framework for defining document lifecycle tracking requirements and retention policies, which meet internal business objectives as well as legal and regulatory requirements for digital records management.

## Elements of the Enterprise Document Repository (EDR)

An effective EDR platform must include the following core capabilities:

1. Document Capture

The system must support direct importation of digital document content and attribute data contained within existing proprietary imaging, document management, workflow, COLD, groupware and similar document processing systems. Optionally, the system should support capture of document metadata only, with the EDR providing a reference to its current storage location.

2. Document Management

The system must be capable of managing any form of digital document, including a variety of image formats, textual and compound documents, electronic files and digital audio/video formats. Users must be able to index and categorize documents, and document storage locations must be referenced. In addition, the system should track document access and retrieval activities.

3. Document Storage

The system must support on-line, near-online and off-line document archiving to multiple storage mediums, including digital formats such as optical disk, CD-ROM, digital video disk (DVD), tape and magnetic disk as well as analog formats, such as microfilm and paper. The system must manage the use of media for document storage, support single or multiple document storage locations, migrate documents between storage media as required by document lifecycle retention requirements, and cache documents to fulfill retrieval requests from system users.

4. Document Access

The system must present a unified search and retrieval paradigm, where documents can be accessed, retrieved and viewed under a consistent usage model regardless of content and storage location. The system should support intuitive search, browsing and navigation capabilities, allowing documents to be organized in a manner consistent with business requirements and accessed with the aid of visual query interfaces with both simple and advanced searching tools.

5. Document Retrieval

The system must be capable of delivering selected documents to the desktop, viewing or otherwise rendering specific documents, and include functions for personal manipulation of the document such as annotation, copying and printing.

6. Document Exchange

The system must provide open access to third-party applications, allowing documents to be extracted, contained by a standard file format and transported as required to support enterprise interoperability objectives.

7. Document Output

The system must support personal and high-speed batch document printing, publication of document collections to CD format, and exportation of repository documents to the network file system for use by third-party applications.

8. Document Disposition

The system must support the removal and purging of documents from the repository based upon a predetermined document retirement policy. In addition, the system should be able to provide an audit trail of document migration and disposition.

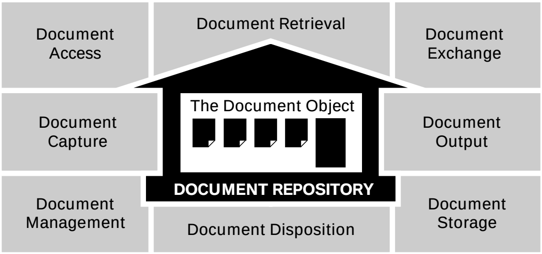


**Figure 2: Enterprise Document Repository Functional Components**

In addition to these capabilities, the EDR must include a set of system management and administration capabilities suited to a standard client/server environment with a clear definition of use and administration. Further, since it is aimed at the enterprise level, the system must support the highest degree of platform scalability as measured by total document storage capacity, retrieval performance and user concurrency, and must provide a highly-reliable and available software service, with data integrity and fault-protection features to maintain continuous system operation.

## The Document Object Model

Central to the EDR and to each of the above functional capabilities is the concept of the “document object”, which is a universal container of document information. Whether it is images, pictures, text, graphics, electronic document files, HTML documents, audio, video, messaging files or workflow process maps, this information is collected as a document object with a unique and persistent identity within the EDR. The container can be conceptualized as either a single document or a folder, which is some business-specific collection of documents.

The document object contains not only the document content, but also all the attribute data that has been linked to a particular document. Referred to as metadata, this attribute data includes index values that have been manually assigned to the document as well as system-assigned attributes, such as creator and creation date.

By nature of its role as “universal container”, the document object must be standards-based so that it is fully interoperable with other business applications and productivity tools. While a definitive standard has not yet emerged, it is likely that the document object will comply with Microsoft’s ActiveX architecture as well as the Content Model as currently defined in the DMA (Document Management Alliance) technical specification.

With document collections, the document object should be capable of either containing or referencing its associated documents. In certain applications with advanced foldering requirements, documents may exist within multiple folders but for performance and capacity reasons are merely referenced by their parent folder.

Other applications with more straightforward single-folder/single-document needs may elect to have all folders physically contain their constituent documents. In either case, the system should allow the folder to be combined into a single file for document exchange purposes.

While each of the eight previously-defined functional capabilities are necessary for the EDR, the two most fundamental and critical services are:

- document management and indexing, to organize and catalog documents in the repository, register their storage locations, manage their access and track their utilization across the enterprise;
- document storage and archive, to support the long-term, secure storage of documents, regardless of their content, as well as provide reliable references to their storage locations.

## Document Management

A variety of EDR document categorization schemes can be employed depending on the particular business requirements. Examples of potential document categorization schemes are included in the table below.

| Hierarchy Level | Option 1 | Option 2 | Option 3 | Option 4 |
|---|---|---|---|---|
| 1 | Repository | File Room | Library | Warehouse |
| 2 | - Collection | - Cabinet | - Catalog | - Room |
| 3 | -- Folder | -- Drawer | -- Section | -- Shelf |
| 4 | --- Sub-folder | --- Folder | --- Document | --- Box |
| 5 | ---- Document | ---- Sub-folder | (unused) | ---- Folder |
| 6 | (unused) | ----- Document | (unused) | ----- Document |

**Table 1: Document Categorization Schemes**

## Document Indexing

The EDR system should, as much as possible, rely on automated indexing of documents as they are captured from any potential source. While some form of interactive input mechanism will be required, system users in general should not need to carry out time-intensive manual indexing activities. Ideally, visual modeling tools can be used to automatically extract attribute field names from the original system source and enter these into the EDR. These index values should include a set of standard fields (e.g., creator, creation date and document title) to help establish a consistent records management discipline as well as an arbitrary set of fields that are customized to business-specific needs.

Since the EDR's value is directly linked to its ability to serve large corporate environments, the EDR index service should be scalable to support high-volume and high-performance requirements. The EDR index should be capable of indexing and maintaining references to 10, 50 or well over 100 million documents, and should provide system access and query support for up to a few thousand concurrent users. In addition, the indexing service must be fault-tolerant, with protection from catastrophic incidents at given sites (e.g., damage from fire, flood, tornado, etc.) as well local infrastructure failures (e.g., server crashes).

## Auditing

The EDR should include document history and utilization tracking functions that provide an audit trail of when documents were committed, archived, migrated, retrieved or otherwise handled. Reports that summarize the results of system auditing and user activity tracking should be generated to describe the following:

- User concurrency and total connect-time for specific users;

- Document access and retrieval activity, sorted by specific document, document category, user or user group;
- Type of searching activity performed; attribute or content-level search; explicit search criteria or otherwise;
- Summary of actions performed by individual users or user groups;
- System performance metrics, including total search and retrieval times;
- Summaries of system resources, including name, location, status and operating histories;

## Access Control

To protect documents from unauthorized access and utilization, the EDR should provide a number of access control and security functions, including:

- Definition of authorized system users, organized by user groups,
- System login and password protection,
- Access control lists for each major document category, assigned by user and user group,
- Access privilege assignments including rights to view the document index, view the document, view the object data, copy the document and print the document.

## Document Storage and Archive

Another critical function of the EDR is to provide a management infrastructure around the long-term secure storage of documents and to provide reliable references to their storage locations. Unlike many currently-available storage solutions that rely on a single device, the EDR must include the ability to store documents of any kind, on a wide range of storage media, including:

- magnetic storage devices and media;
- optical drives and jukeboxes, including WORM 5¼-, 12- and 14-inch media;
- 5¼-inch rewritable media;
- CD drives, transporters and jukeboxes, including standard CD, CD-R and emerging DVD formats;
- digital tape drives and autochangers;
- micrographic imaging devices and roll film.

The storage service should support the notions of on-line, near-online and off-line storage, generally defined as follows:

- On-Line: Where an archive or retrieval request can immediately be serviced by the storage system, for example with magnetic media or when an optical platter is mounted and under the drive-head.
- Near-Online: Where some intermediate media manipulation is required by a robotic mechanism (jukebox or autochanger) before the request can be serviced.
- Off-Line: Where a manual intervention is required to mount the archive media before the request can be serviced.

Regardless of its content or lifecycle stage, EDR documents must be available to all users from any network location. The document's physical storage location is irrelevant to the user's retrieval process or needs for access.

## Document Archive

The storage service should automate the entire document archive process, including the ability to define retention schedules that determine storage duration, media migration and disposition policies for repository documents. The software should also enable control of how document collections are physically organized and assigned to the archive media, including support for archive storage groupings (“archive profiles” or “archive classes”) consisting of some set of related archive media (i.e., one or more optical volumes, tape volumes, etc.). Other features should include the ability to manage storage allocations to archive media, such as reserving space on optical platters. The use of self-describing archive media will enable document location data to be restored from the media itself.

Just as for indexing services, the EDR archive storage service should be capable of supporting high-volume, high-performance levels, and should be designed to ensure high reliability and protection from data loss. Typical EDR systems should be capable of storing and maintaining reliable references to 10, 50 or well over 100 million documents. In addition, high-volume archive throughputs ranging from 10,000 to over 25,000 documents per hour should be achievable. Finally, the EDR should be capable of delivering page-oriented documents (first page) in less than 2 or 3 seconds from magnetic media, less than 5 to 7 seconds from mounted optical media, and less than 20 to 30 seconds from unmounted optical media.

### Intelligent Caching

The storage system should provide intelligent caching facilities that automatically cache archived documents to an optimal position local to the requesting user. The system should also support pre-fetching operations, where a retrieval request can be scheduled with the storage service so the document will exist in archive cache in advance of the actual user need.

### Document Disposition

Just as a retention schedule should control the archive and migration of documents among various media, a document disposition schedule should dictate the retirement of documents from the repository. The retirement and possible destruction of documents should be based on the document’s age, corporate value, and any legal retention requirements.

When determining document disposition policies, it is important to consider the legal status of the variety of media used in the document repository. As with paper documents, the legality of micrographic media is well established. Numerous federal and state statutes ensure the legality of micrographic media, microfilm, and microfiche, for recordkeeping purposes and as evidence in trials or other legal proceedings. The legal status of electronic document images and electronic documents is less established. In the absence of federal laws governing the legal acceptance of electronic documents and document images, it is critical to ensure that electronic documents have integrity, that is that they are true representations of original paper documents and that they have not been inadvertently or fraudulently altered or destroyed. Clearly articulated retention and disposition schedules, as well as measures that encourage and demonstrate that these policies have been correctly and consistently implemented, protect the integrity of the document repository, as well as documents that have been removed from the depository.

### Document Capture

Document capture encompasses all functions associated with the sourcing, transformation and entry of digital documents into the repository. Most of the documents imported to an EDR are likely to exist in some electronic form within proprietary document processing systems. That means the primary capture model will involve extracting these documents at the source and transforming them into a format suitable for entry. Optionally, the EDR should include the ability to reference documents within existing third-party repositories, including imaging, document management, workflow, COLD and groupware systems. Whether it is complete importation or "referencing in-place," each potential document source will require the existence of specialized content and metadata capture gateways that can connect to the source and perform filtering and transformation as necessary to create an EDR-compatible input stream. The specific activities to be performed at each gateway include:

- Source System Connectivity: Establishing and maintaining a network connection to the existing document source using whatever access mechanism and networking protocol is supported by the source system.
- Content Importation and Conversion: Isolating the type and format of the content within the source system and performing format conversions as necessary for importation to the EDR.
- Metadata Mapping and Filtering: Extracting or referencing the document attribute information from the source system, filtering it as required, and mapping it to specific fields using the EDR indexing services.
- Document Containment: Optionally collecting both the document content and attribute data and placing them within a standard object container.
- Repository and Archive Committal: Staging the input stream of source documents and invoking repository "commit" and "archive" commands to place them under EDR system control.

As mentioned earlier, manual indexing activities should be minimized; instead automated indexing methods and attribute mapping should be emphasized. In addition to capturing documents from third party systems, the EDR must be capable of capturing documents that have been written to CD-ROM media, documents that have been written to a network file system, and finally, documents that are currently paper-based.

## Document Access and Retrieval

Document access encompasses all user activities associated with browsing and searching the repository for documents of interest, including navigation across different logical document categories and storage containers, executing queries against the repository index to locate specific documents, and examining document attribute information that has been returned based on the query. Assuming the user chooses to support in-place referencing, this will require coordinated query across multiple, independent repositories. Document retrieval includes those functions required to support direct viewing of documents stored within the EDR. The client interface, which allows users to perform both of these functions, should be intuitive and user-friendly so that users with a range of computer familiarity and literacy are able to navigate and retrieve information.

## Document Browsing and Navigation

Users must be able to easily and intuitively browse and navigate among the document categories present in the EDR. An ideal browser would resemble the Microsoft Explorer interface, using a tree diagram that can be expanded and collapsed using the mouse or keyboard controls. A multi-panel interface enables the

EDR to display document categories in one panel and the attributes corresponding to the selected category in another panel. The design of this interface should allow customization based on user preferences or requirements.

Users should be able to perform various searches against the repository index to locate specific documents, including the ability to:

- Search for documents within a specific category using structured attributes only, using full-text attributes only, or using a combination of structured and full-text attributes;
- Search across multiple document categories using any combination of structured and full-text attributes;
- Search the entire repository (i.e., global search) using relevant combinations of structured and full-text attributes;
- Search using either explicit or wild card criteria with support for multi-field queries using AND, OR, NOT, and similar search operators. In the case of full-text queries, this includes support for fuzzy search capabilities that relaxes the requirement for an exact word match.

Assuming the executed query retrieves an unsatisfactory set of results, the system should provide a mechanism to narrow or expand the search through the provision of additional criteria. The system should enable the naming and saving of repository search operations to allow them to be reused for personal or corporate-wide purposes.

The ideal EDR product will offer a more intuitive, visual query interface than the typical field and string search capability found in most document management systems. In the best case, users should be guided through the search process using some combination of visual cues (presented as context-sensitive interface icons and usage guidelines) or “search wizards” that provide step-by-step search instructions.

## Document Retrieval, Viewing and Manipulation

Document retrieval is the companion feature set to document access and makes up the other major component of the EDR client interface software. While document access focuses on locating items of interest, document retrieval initiates reading of the document from archive storage and supports rendering it at the user desktop.

The EDR system should support viewing of repository documents through use of standard off-the-shelf viewing software. This can be either stand-alone executables that launch as separate windows upon retrieval, or ActiveX components that support in-place viewing as a seamless part of a multi-panel interface. The system should provide tools for associating viewer software with the document type or representation being delivered. This can range from using a general purpose viewer (e.g., the Microsoft/Eastman Software Image 95 viewer for bitonal and color images) or a Web browser (e.g., Netscape or Microsoft Internet Explorer) to launching native applications for viewing purposes (e.g., launching Microsoft Word to view a Word file).

In addition to viewing documents, the EDR retrieval software should provide an interface for viewing folderized collections of documents. At a minimum, the system should support the presentation of a folder table-of-contents that lists the documents contained within that folder.

Since the EDR system supports document archival to on-line, near-online and off-line storage media, the EDR should provide users with information about how long it will take to retrieve information based on where it is stored. In addition, the system must notify the user if manual loading of the media is required to access the requested information.

While native annotation capability is not required in the EDR, certain viewing components employed by the customer may enable annotations as part of their standard feature set. These viewers typically provide simple markup tools such as graphics (e.g., line, oval, rectangle), post-it notes, yellow highlighting and textual overlays. These annotation tools support enhanced document communications (e.g., reviewing document content) between users over the network. Given the nature and purpose of EDR, any such annotations are made on working copies of the document and do not effect permanently-stored repository records.

In addition to document annotation, the system should provide users with local document manipulation capabilities, including:

- copying the document or folder to a file,
- attaching a document or folder to an e-mail message,
- printing a document or folder to a local or network printer,
- faxing a document or folder through a local or server-based fax gateway.

## Document Exchange

As an enterprise resource that addresses cross-functional business requirements, the EDR must allow other applications to access the document information contained in the repository. This includes support for query and retrieval by third-party library systems, extraction of documents for business or personal use, transporting documents between enterprise locations using the Internet or other messaging services, and moving documents between two or more operationally-independent EDR systems.

The system should support the exchange of EDR documents with messaging and groupware applications operating over public and private networks. This should include support for the following delivery environments:

- Microsoft Exchange and other MAPI-compliant messaging systems,
- Internet-based electronic mail packages,
- Lotus Notes/Domino databases and applications.

Further, the EDR system must support interoperability with third-party library services being used for active lifecycle document management purposes. Specifically, the system should be compliant with the two major document management standards initiatives currently evolving in the marketplace:

- The Open Document Management API (ODMA) standard for client-side integration,
- The Document Management Alliance (DMA) specification for server-side integration.

Finally, the system should support document transport between two operationally-independent EDR platforms over public and private networks. This capability is intended for use in both inter- and intra-enterprise document exchange scenarios. Between enterprises, the system at one corporate location must support batch extraction of documents that can then be imported to a system at another corporate location. These two systems are by definition distinct and share no server resources whatsoever. Within enterprises, multiple EDR platforms may in certain circumstances be configured as separate operational systems to avoid dependence on or contention for shared resources. In either case, the system should support exchange of one or more documents using standard store-and-forward messaging infrastructures as the transport mechanism.

### Document Output

This functional component of EDR provides services for outputting documents in various formats, including:

- personal and batch printing operations, including the ability to print both document content as well as the metadata, as appropriate. Supported printers should include any compatible Windows printer as well as higher-speed (30-60 ppm) printers available from vendors like Xerox or QMS.
- local and network fax distribution of one or more documents through standard facsimile gateways.
- extraction and "publication" of EDR documents to CD format, including all document content and metadata, as well as an optional document viewer that can be used to retrieve and render the documents written on the CD.
- exportation of EDR documents to the network file system where they can be imported and utilized by other enterprise applications.

### EDR System Management Requirements

In addition to the specific functional components for capturing, managing, storing and exchanging documents, the EDR must include system management capabilities that ensure ease of installation, administration and maintenance in a client/server environment.

Visual tools should be included to make EDR system configuration a straightforward process. Basic system installation and resource set-up features should include:

- Program file installations and updates
- Server naming and configuration
- System sizing parameters (e.g., high-water marks)
- Network access paths and volume assignments
- User security and access control lists
- Document retrieval preference settings
- Archive media selection and collection assignments
- Archive retention policies and rules of migration
- Cache storage allocations and server associations
- Resource logging and control parameters
- System backup strategies

A robust backup system that recognizes the scale of the EDR system should generate a reliable backup in a minimum amount of time. This should include the ability to perform "hot" backups that can be performed incrementally and while the system is running.

## EDR Architectural Considerations

The EDR can be designed as an integrated network service, comprised of a few cooperating components that fit within standard client/server infrastructures (see Figure 3). Potential EDR components include:

- Index Server: The server component that manages document indexing information as well as related object, system, resource and user-specific metadata. In order to fulfill required scalability parameters, the system should support configuration of the Index Server on either single or multiple servers with partitioned index (document) databases. Even in the case of multiple physical servers, end-users should continue to perceive only one logical system. The system should transparently support search operations across multiple indexing servers and deliver a unified result set to requesting users.

- Archive Server: The server component that manages document location and storage management information. In order to scale as required, the storage system should support configuration on either single or multiple Archive Servers interfacing to single or multiple archive storage subsystems (i.e., the component that interfaces to the physical storage device).
- Archive Subsystem: The server component or process that interfaces to physical storage devices to store (write) and retrieve (read) from the archive media. In the simplest configuration, the storage service and the archive subsystem operate as independent processes on the same server. In other configurations, a single logical storage service can be configured as a set of physically distributed but cooperating storage servers and archive subsystems.
- Retrieval Clients: The client components that transparently access system servers and subsystems to review and retrieve documents.

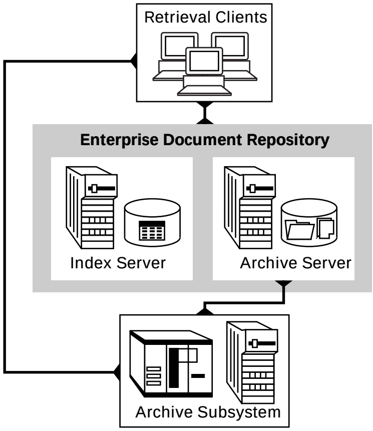


**Figure 3: Enterprise Document Repository Major Platform Components**

## Client Environments

The system should support a standard 32-bit Windows desktop, including support for both Windows 3.x, Windows 95 and Windows NT workstation environments. For system access over the Internet, the system should support predominant web browser products on the market, such as Netscape and Microsoft's Internet Explorer.

For either standard desktop or web-based configurations, the client component should have local disk space adequate for caching document pages for optimized view performance. This will require a minimum of 100 MBytes of storage.

## Server Environments

As a distributed client/server platform, the EDR system should run on either Windows NT and UNIX (Solaris, AIX and HP/UX derivatives). The choice depends on in-house preferences and stated corporate architectural strategies.

For the Windows NT platform, the system should be configured as a true NT service and support the standard NT registry and native NT capabilities for event logging and performance monitoring (through custom NT performance counters). The system should also make management information available through a custom Management Information Base (MIB) for use by SNMP-compatible network management consoles. The system should provide equivalent administration and control capabilities on UNIX and all other platforms proposed for EDR.

Both the Index Server and Archive Server are based in part on standard relational database technology. To the greatest extent possible, the system should mirror database platforms already in use so that in-house database development and administrative resources are available. This is particularly true for the Index Server, which manages customer-specific document metadata that will likely require access by other enterprise applications. In general, the supported database platforms should include Microsoft SQL Server, Oracle, Sybase, Informix and DB/2.

As with the client platform, the EDR system should support Netscape's Enterprise Server and Microsoft's Internet Information Server for Internet and Intranet-based access to repository documents.

### Network Environments

EDR should coexist with and, where appropriate (e.g., for distributed caching), leverage the services of the existing enterprise network. The system should support protocol-independent network communications between client and server components. This should include primary support for communicating over Internet-standard TCP/IP, with optional support for IPX/SPX, NetBIOS and other network transports. With the exception of performance sensitivity, the system should operate transparently on whatever physical network is available; either local or wide area, public or private. This should include support for Ethernet, Fast Ethernet, Token Ring and ATM connectivity, as well as ISDN and dial-in access at 28.8 kbps.

## Guidelines for Implementing the EDR

If you believe your company could benefit from the creation of a repository for managing the long-term storage of digital documents from multiple systems, here are ten steps that can help get your project started:

1. Perform a corporate-wide inventory of existing automation systems that are producing, managing and storing digital documents within your company. Specifically identify the functional and content categories of documents, the volumes being managed, and the primary department or business function the system was designed to support.
2. Establish end-user needs for cross-functional access and retrieval of the digital documents contained within the proprietary document stores identified above. Attempt to quantify both the source and frequency of access as the document collection ages over time. Use this information to determine which document stores should be aggregated, and to set priorities for archiving documents to the repository once you are ready to begin deployment.
3. Review each major category of digital document to define the retention requirements throughout the document lifecycle. Specifically focus on the post-archive phase of the lifecycle to establish the overall retention period as required by corporate policy and/or industry regulation.
4. Extend the scope of your document inventory analysis to include documents not yet addressed by an automated system. Documents that have minimal processing requirements could bypass other automated systems and be delivered directly to the EDR. Perform similar content, volumetric and cross-function utilization studies as described in steps 1-3 above.
5. Assuming evidence of real user need, proceed to analyze the metadata or attributes used to index each of the document categories within the document stores previously identified. Isolate those attributes that can be used as primary index fields for optimal search and retrieval operations against the enterprise repository. Use this analysis to resolve metadata field naming and data type inconsistencies as necessary.

6. Perform a more detailed review of each of the operational document processing systems identified above to establish specific criteria for determining when documents should be off-loaded or archived from the existing system to the enterprise repository. This may be based upon process-specific, volumetric or aging parameters depending on the orientation of each system and the nature of the document information it stores.
7. Develop an archive media strategy for all digital documents to be managed by the enterprise repository with special consideration of any industry-specific regulations that may govern the legality of particular media types for permanent document storage. Also consider any limitations that particular content types may impose on media selection to determine if content conversion is required before committal to the enterprise repository.
8. Determine if these is a cost-justifiable business need, based upon defined retention policies and overall patterns of access, to migrate documents between different types of archive media (e.g., from optical media to film) in the latter stages of the document lifecycle. This should include consideration of policies for retiring documents from the repository when there is no future need for access.
9. Map the functional business organization and the projected hub(s) of user retrieval activity against the existing technology infrastructure to establish needs for distributing repository components throughout the enterprise. This analysis should focus on requirements for optimizing retrieval efficiency by locating archive storage resources in proximity to the largest potential community of requesting users.
10. Using the insights gained from the steps above, develop a unified document indexing strategy for the enterprise document repository. The objective is to define an overall access framework for repository users that balances logical business requirements for document organization with physical infrastructure limitations in network, database and storage technologies.

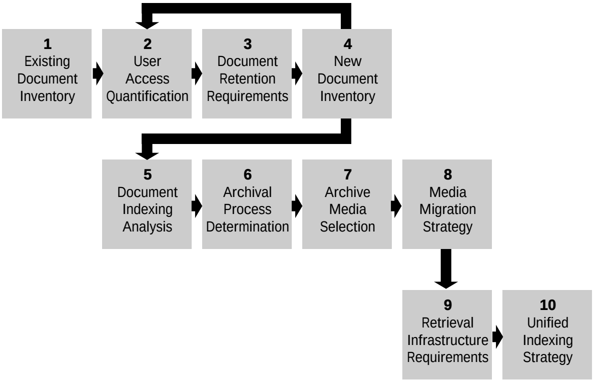

1
Existing Document Inventory
2
User Access Quantification
3
Document Retention Requirements
4
New Document Inventory
5
Document Indexing Analysis
6
Archival Process Determination
7
Archive Media Selection
8
Media Migration Strategy
9
Retrieval Infrastructure Requirements
10
Unified Indexing Strategy

**Figure 4: Document Warehousing Needs Assessment Process**

## Summary

With the above activities completed, an organization is now ready to proceed with a document warehousing project. The goal of the document warehousing project is not to replace or obsolete the disparate, unstructured management systems already in place, but rather to aggregate all, or a portion, of these individual corporate information assets into an easy-to-access and reliable information knowledge base: the EDR.

The EDR has a fundamentally different orientation than current-generation document management systems. While document management system focus principally on document processing activities relevant only during the most active phases of the document lifecycle, such as revision control, check-in/check-out and configuration management, the EDR is specifically designed to fill the void that exists in the latter stages of this lifecycle. In these stages, the focus turns to enterprise accessibility, long-term storage and information asset protection. As such, EDR is primarily oriented to provide a universal store for unstructured document information with an emphasis on access consistency, multimedia management, intelligent linkage to structured data and secure, high-capacity storage. In this context, it is more of a digital records system than a document management technology.

By building an enterprise document repository, an organization can improve customer service, streamline operations and achieve a competitive advantage. What data warehousing has done for structured data, document warehousing can do for business documents and other unstructured information. The combination of these initiatives provides organizations with a unique ability to transform disparate information bases into comprehensive business intelligence. To the corporate business user, that means the difference between information overload and information empowerment.

# Customer Profile: Document Warehousing—A Framework for Understanding

Debra knows next to nothing about intelligent forms recognition, legacy application interfaces, object modeling, device integration or unstructured data.

But as one of the most conscientious and capable claims processors for a leading property and casualty insurance carrier, she knows a lot about the importance of having immediate access to the right information, when and where she needs it. She also realizes that currently, it is often next to impossible to locate and link stored information residing in corporate databases with associated source documents existing on paper, microfilm or electronic media.

In this document management environment, Debra will spend more and more of her time analyzing information from a wide variety of sources, and much less time collecting needed data and digital documents in a useable form.

The potential impact on basic claims processing, not to mention the global environment for information and image management, will be profound.

On this particular morning, Debra is attempting to reconcile a disputed claim stemming from an apparent lapse in the claimant's insurance coverage. Without ever leaving her workstation, she compares information extracted from the company's master client database with microfilmed and scanned images of the client's original application and unpaid invoices, respectively.

Debra also has the opportunity to review digital images of the claimant's loss supported by an audio-appended description of the storm-damaged equipment. After calling up document images of recent correspondence with the customer, Debra supplements selected files with her personal recommendations that she electronically transfers to her supervisor's desktop. She also e-mails the same files to the carrier's legal counsel in New York City.

In a matter of minutes, Debra has successfully completed her investigation, never giving a second thought to the initial methods of data and source document capture which occurred over a long period of time; the applications they were created in; the indexing software which link associated files; the interfaces which bridge disparate applications; the system security protocol that prevents unauthorized access and protects information integrity; the gateways and drivers that permit information distribution down the hall or across the country; and the software and hardware environment in which the information resides.

Instead, Debra can focus her energies and skills to tackle the job at hand-to maximize her contributions to the organization she serves. Helping people and companies work smarter and faster…that's what document warehousing is all about.

## Document Warehousing—Kodak's Participation

Document warehousing represents a powerful new strategy for unifying the management and storage of unstructured information for a broad range of business applications. Leveraging interoperability standards such as DMA and Kodak's core competencies in digital capture, image and information management, mixed-media storage and archival science, document warehousing is well positioned to become the business solution that finally fulfills the technological promise of increased productivity and profitability.

(continued from page 1)

## Kodak Business Imaging Systems: Enabling Better Document Access

**Systems and Solutions: Pumping up Processing.** Beginning in the 1980s, Kodak developed a series of computer-based document management systems which continue to evolve. Presently, organizations can re-engineer their work processes for better control and efficiency with *Kodak Digital Science*™ Enterprise Imaging System (EIS) software. In mission-critical applications such as health claims processing, EIS automates workflow, cutting hours—or even days—out of cycle times. Kodak also offers a variety of other systems for efficiently filing and retrieving document images stored on electronic and film media. Typically, these systems are used to resolve exceptions or to reference documents in response to customers or internal inquiries, when speed and access are crucial.

**Service: Support from Plan to Help Desk.** To assist customers implementing imaging solutions, Kodak offers a suite of professional services that include feasibility analysis, installation, conversion, and training. Kodak also supports customers indirectly by providing technical tools and consultation to a third-party network of nearly 100 data and document conversion service bureaus.

### A 70-year head start on tomorrow’s solutions

Within Kodak, Business Imaging Systems remains closely aligned to the corporation’s strategic goal of being the world leader in imaging. The group works with other Kodak divisions, and with industry leaders such as IBM, Microsoft and Eastman Software through Kodak’s strategic business alliances. As a result, Business Imaging Systems can access a huge array of resources, expertise and technology to address a customer’s particular document imaging problem.

## Selected Product Listing

**Media**
*Kodak Imagelink* Microfilms
*Kodak* COM Microfilms
*Kodak* Writable CD Media with *Infoguard* Protection System
*Kodak* Digital Archive Media
*Kodak* 14” Optical Disk Media

**Hardware**
*Kodak Imagelink* Desktop Microfilmer III
*Kodak Imagelink* Microimagers 30 and 70
*Kodak Imagelink* Intelligent Retrieval Workstation 1000
*Kodak Imagelink* Digital Workstation 2000
*Kodak Digital Science* Scanner 5500
*Kodak Digital Science* Scanner 7500
*Kodak Digital Science* Scanner 9500
*Kodak Digital Science* Scanner/Microimager 990
*Kodak Digital Science* Capture Subsystem
*Kodak Digital Science* Document Archive Writer, Model 4800
*Kodak Digital Science* PCD Writer 600
*Kodak Digital Science* Disc Transporter
*Kodak Digital Science* CD Automated Disk Library 100 and 150
*Kodak Digital Science* Optical Disk System 2000 and 2000E
*Kodak Optistar* Image Writer

**Systems/Solutions**
*Kodak Imagelink* Business Solutions / PC Plus
*Kodak Imagelink* Application Services for PC LANS
*Kodak Imagelink* Application Services for CICS, IMS
*Kodak Digital Science* Digital Document Archive System
*Kodak Digital Science* Data Management System
*Kodak Digital Science* Mainframe COLD System
*Kodak Digital Science* Writable CD COLD System
*Kodak Digital Science* MultiStore Software
*Kodak Digital Science* CD File Store Solution
*Kodak Digital Science* Professional Capture Management Systems

**Services**
*Kodak* Professional Services ~ Feasibility and Qualification Studies, Installation and Post-Sales Support, Training, Implementation Services
*Kodak* Microfilm Support Programs ~ Disaster Recovery, Quality Control, Environmental and Safety
*Kodak* Customer Equipment Services

(Note: Kodak, Imagelink, Digital Science, Infoguard and Optistar are trademarks.)

Kodak

EASTMAN KODAK COMPANY
Business Imaging Systems
Corporate Headquarters: Rochester, New York 14653-6304
(800) 243-8811

In other areas contact the Kodak company serving your country.

Visit our home page on the World Wide Web at
http://www.kodak.com

This Technology Guide is one in a comprehensive collection of Guides that supply objective information on technologies related to the Internet, Communications & Networking, Data Warehousing, and Document Management. Each guide presents complex technology concepts in practical and understandable terms to assist you in your education, evaluation, and decision-making processes.  Visit our Web Site to view and print this Guide, as well as all other Guides.

www.techguide.com

A-5412 • CAT: 888 8711

Produced and Published by

ATG

The Applied Technologies Group

One Apple Hill, Suite 216, Natick, MA 01760
Tel: (508) 651-1155  Fax: (508) 651-1171  E-mail: info@techguide.com